The Epistemology of Equivalence: Symmetries and Dualities

Dominik Ehrenfels

Wolfson College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2025

Abstract

This thesis addresses the question 'Under what circumstances is it warranted to interpret symmetry-related models as equivalent?'. Moller-Nielsen [2017] distinguishes between interpretationalist and motivationalist attitudes towards this question. Interpretationalists maintain that observationally indistinguishable symmetry-related models may invariably be interpreted as equivalent. Motivationalists on the other hand argue that one is only licenced to do so once one has found a coherent account of the ontology shared by the symmetry-related models. This account is to be supplied in the form of a core model common to the symmetry-related models. This thesis contributes to the debate by defending a moderate form of motivationalism. It sets out by arguing for an approach to theory interpretation this thesis calls ameliorative interpretation. Subsequently, a moderate motivationalism is defended against both interpretationalism and more demanding forms of motivationalism. The motivationalism defended here is distinguished by (1) insistence that the construction of common core models is crucial to judgments of equivalence for symmetry-related models, (2) the view that where a common core construction falls short of providing a coherent ontology for the symmetry-related models, a judgment of equivalence can in certain cases nonetheless be secure, and (3) rejection of demands to the effect that the common core must take the form of an intrinsic theory in the sense of Jacobs [2022a]. The thesis also addresses how common core theories are to be constructed. An ecumenicism about the approaches of reduction and sophistication (Dewar [2019]) is defended, and it is argued that reduction is a more promising approach to reformulation than is often assumed. Finally, the thesis considers extensions of motivationalism to (a) nomic possibility and (b) dual theories. With respect to the latter, this thesis aims to show that a phenomenon Putnam [1977, 1980, 1987, 2001] calls 'conceptual relativity' does not pose a problem for motivationalism.

This thesis is approximately 67,500 words long.

Acknowledgments

Funding for this thesis was provided by the Institute for Ethics in AI in the form of a two-year scholarship. I thank them for the generous support.

Adam Caulton was the primary supervisor for this thesis. I will remain forever indebted to him, for his kindness, patience, and generosity, and for seeing something in me when I couldn't. Our long discussions over the years have shaped every page of this thesis, and I will always remember them fondly.

I am also deeply grateful to my secondary supervisor, Alexander Paseau, for his trust and support. Alex gave me the opportunity to teach at Wadham and helped me tremendously with job applications. Without his help and advice, an academic job would have been all the more difficult to secure.

Special thanks are owed to James Read, for taking much time to read my work, give me valuable comments, and meet with me for discussions. I learned a hundred new things every time I stepped into his office.

I'd also like to thank Tushar Menon, Niels Martens, Eleanor March, Neil Dewar, Caspar Jacobs, and the audience at the Oxford Philosophy of Physics Seminar for challenging discussion and insightful comments. Two anonymous reviewers for the British Journal for the Philosophy of Science provided valuable feedback on one chapter of this thesis, for which I am tremendously grateful.

This thesis very much owes its existence also to my friends and family.

I thank Daniel Kodsi, for showing me what it is to be a good philosopher, and what it is to be a good friend.

I thank Nicole An for placing her faith in me and giving her all to support me. Her family, too, has done more for me than I can ever repay.

Finally, and above all, I thank my parents, Thomas and Friederike, my brother Pascal, my aunt Christel, and my grand uncle Herbert, for building me up, for having my back, and for making sure I never had to worry about anything other than finishing this thesis. What I have accomplished, I have accomplished through them.

Table of Contents

1 Introduction

Remarkably, in science we often figure out the maths long before we figure out what the maths is about. It is one thing to find the equations that capture the phenomena. It's another thing to see through the equations and understand what they tell us about the world. That is why even after a theory has been conceived, shaped, simplified, freed of inconsistencies, and given a rigorous mathematical definition, more is to be done: the theory has to be interpreted. By far the best and most notorious illustration of this is quantum mechanics of course, a theory that more than any other has resisted telling us what the world is like.[1] But also in other theories, it is not straightforward to say what the mathematics describes. Take Maxwell's equations for instance: do they model excitations of an ether, or do they describe propagating fields in spacetime? Long after the equations had been put into their canonical form, the question of their interpretation remained.

One of the most important guides to the interpretation of a theory is symmetries.[2] Roughly speaking, a theory has a symmetry whenever it allows for distinct histories that agree on the values of any quantities that by the lights of the theory's dynamics are in principle measurable. In Newtonian Mechanics for instance, one finds histories that differ over the absolute velocities of the bodies in the universe but agree on everything else. When a theory has a symmetry then, there is some leeway for altering the quantities it posits without effecting an observable difference. A symmetry transformation is precisely a variation of quantities along these degrees of freedom afforded by the theory.

Symmetries are widely taken to indicate that there is a mismatch between the quantities described by the theory and the quantities that are in fact physical. After all, norms of good theorizing would demand that we only posit quantities variation of which at least sometimes makes a dynamical difference. It is therefore often assumed that when a theory has symmetries, we can infer that only a subset of the quantities it purports to describe are physical, and that any two models of the theory that agree on the values of all of these quantities are to be regarded as equivalent. Wallace captures these assumptions with the following two theses:

[1] For an overview of interpretations of quantum mechanics, see Wallace [2008]

[2] For discussion of the relevance of symmetries to interpretation, see *inter alia* Ismael and van Fraassen [2002], other essays in Brading and Castellani [2002], Caulton [2015] and Dewar [2019]

The Representational Equivalence Thesis: Given a family of models of a theory which are related by a symmetry transformation, insofar as one model successfully represents a system, so do all the others.

The Surplus Structure Thesis: Given a theory with a symmetry transformation, insofar as the symmetry falls short of being an automorphism of the mathematical structures used to define the theory's models, this points to aspects of that structure which are redundant, do no representational work, and can be removed from the theory without loss.

Wallace (2022b, p. 328)

But can we immediately conclude from the fact that a theory has symmetries that it gets things wrong? That is the central question this thesis is intended to address. One popular view, called interpretationalism in the literature, answers yes. Interpretationalism is endorsed in some form or other by Saunders [2003], Huggett [2011], Dewar [2019], Wallace [2022a] and Luc [2023]. Saunders [2003] additionally advances an invariance principle, according to which at most the structure invariant under symmetry transformations may be assumed to be real.

Another view, known as motivationalism, urges caution (Moller-Nielsen [2017]). Motivationalists agree that symmetry-variant structure is to be viewed with suspicion. It motivates us to replace the theory with one in which the variant structure is excised – hence the label 'motivationalism'. However, its proponents point out that interpreting symmetry-related models as equivalent straightaway comes at a cost. Theories offer us an account of what the world is like; what its constituents are and what its fundamental structure is. It is by appeal to these entities and this structure that theories purport to explain the phenomena they predict. But to say that where the theory sees a difference, in reality there isn't one, as interpretationalists do, is to deprive us of that account and of the explanations based on it (Moller-Nielsen [2017], Read and Moller-Nielsen [2020a], Martens and Read [2020], Jacobs [2022a]). We are left with a black box: a theory that gets the predictions right but beyond that is not to be trusted. This only is a reasonable view to take if we can be certain that a better theory – a theory with the same explanatory power but without the symmetry-variant structure – is out there to be found. Motivationalists question whether such certainty can be had without proof. They argue that we are only warranted in interpreting models related by a symmetry as equivalent once we have replaced the old theory by a new one in which we no longer find symmetry-variant structure. Instead of a black box, we then have a theory that provides us with an ontology and a dynamics that can provide explanations of the phenomena the theory describes.

To see why motivationalists think that finding such a new theory must precede judgments of equivalence, it helps to look at the example of Newtonian Gravitation Theory and its history.[3] In Newton's theory of gravitation, it is assumed that there is an absolute space, enduring through all time. This leads to a problem, since the dynamics of the theory is not sensitive to variations of absolute velocity, i.e. velocity relative to absolute space: uniformly adding a constant velocity to the motion of all the bodies does not affect the relative motions and hence does not result in an observationally distinguishable history, and so there is no way of telling whether an object remains at the same point in space throughout or is moving inertially. The theory has a symmetry. Nonetheless, according to Newton's theory, there is a fact of the matter as to what the velocity of each object is relative to absolute space. The absolute velocities of the particles are therefore a quantity that is in principle undetectable. One would suspect therefore that attributions of absolute velocity do not track anything in the world. Absolute velocities are surplus. This is the verdict interpretationalists are happy to settle on as soon as the symmetry of velocity boosts is detected. They are certain that Newton's account of the world is wrong and that a better account is out there. But once one appreciates the history of the search for such an account, the interpretationalist position may begin to look epistemically irresponsible.

For the longest time (for centuries!) a coherent account of what the world could be like so as to allow for Newtonian dynamics even though objects do not possess absolute velocities remained elusive. Relationalists such as Leibniz offered an ontology that excluded absolute velocities: according to them, the world was nothing but a sequence of arrangements of bodies relative to one another. All that was real were the relative distances between the bodies at a time.[4] Much philosophical argument was offered in favour of this ontology, but what relationalists could not do was show how to construct a theory from this ontology that replicates the predictions of Newton's theory. Indeed, the task seemed impossible: Newton's bucket argument seemed to show that the relationalist's ontology is too thin to draw all of the distinctions needed to replicate Newtonian dynamics – something must be missing from the relationalist picture.[5]

Only much later did we learn how to make sense of a world with Newtonian motions but without Newtonian absolute velocities. To get there however, tremendous mathematical

[3] See also discussion in Dasgupta [2016]
[4] The *locus classicus* here is the Leibniz-Clarke correspondence (see Alexander [1956])
[5] For a survey of the history of relationalism, see Pooley [2013]

and conceptual breakthroughs were necessary. The key was to think of space and time as inextricably woven together in what we now call spacetime. With this idea in hand, one could define a spacetime that carries Galilean structure. Bodies in Galilean spacetime possess relative but not absolute velocities. So, finally, we had what motivationalists are after, an account of what the world might look like that explains how absolute velocities in Newton's theory can be viewed as mere gauge.

Motivationalism was first put on the map by Moller-Nielsen [2017]. It has subsequently been defended *inter alia* by Martens, Read, and Jacobs (Martens and Read [2020], Read and Moller-Nielsen [2020a, b], Jacobs [2022a]). Others however consider it misguided or overly demanding. Some of Dewar's work rests implicitly on a rejection of motivationalism (Dewar [2019], [2023a]). Interpretationalism has also been more directly defended by e.g. Wallace [2022a] and Luc [2023]. In this thesis, I intend to investigate the disagreement in detail.

The aim of this thesis is to develop and defend a broadly motivationalist attitude towards the interpretation of theories. But before doing so, it will be good to take a step back and ask: what exactly is it to interpret a theory in the first place? Chapter 2 aims to shed some light on this question. I distinguish a number of distinct notions of interpretation and show how they can help clarify the debate between motivationalists and interpretationalists.

Chapter 3 then is dedicated to an examination of motivationalism. I begin by defending motivationalism against a number of objections. Here, I address *inter alia* pro-interpretationalist arguments based on the dynamic approach to spacetime structure and what one might call "easy road" approaches to the elimination of symmetries, principal among which is Dewar's [2019] "external sophistication". Having made the case for motivationalism, I then turn to the nuances of the view with the aim of identifying the best, most defensible version of motivationalism. I particularly emphasize the importance of reformulation: according to the view I defend, we are only permitted to interpret symmetry-related models as equivalent once we have reformulated the theory in which they occur in a way that eliminates the structure variant between the models. This puts me in opposition to motivationalists who maintain that it is already legitimate to interpret symmetry-variant models as equivalent once a candidate ontology for a theory without the variant structure has been identified, even when such a theory has not yet been constructed.

I further distinguish my own version of motivationalism by arguing against Jacobs' [2022] view that reformulations that are apt to warrant a judgment of equivalence for symmetry-variant models must meet what he calls a requirement of *intrinsicality*. An intrinsic theory is one in which all structure putatively directly represents physical structure. I argue that the ideal of intrinsicality is largely unattainable and that knowledge of equivalence for symmetry-variant models can be grounded also in non-intrinsic theories.

Chapter 4 picks up the thread of reformulation. Having argued in Chapter 3 that reformulation is crucial to the interpretation of theories with symmetries, I now consider what form such reformulation ought to take. Dewar [2019] distinguishes two approaches to reformulation, which he calls reduction and sophistication. I argue that while reduction has recently been the target of much criticism, it is much more resilient and fruitful a strategy for reformulation than is often appreciated. This is due to an overly restrictive conception of what reduction can look like, I claim. I demonstrate that reduction is a viable strategy for reformulation by showing how to reduce Newtonian Gravitation Theory with respect to its scale symmetry. This reduction is based on Barbour and Bertotti's relationalist theory of motion.

Chapter 5 is in many ways an addendum to Chapter 4: one thing that emerges from the discussion of Chapter 4 is that reformulation can lead to the elimination as unphysical not just of symmetry-variant structure but of entire models of the original theory: in the transition from NGT to Barbour Bertotti Theory for instance, models of NGT with non-vanishing total angular momentum are not recovered. This suggests a motivationalist attitude towards the question of which of a theory's models one ought to interpret as physical. For in many cases, our theories have strange, seemingly pathological models that we might be inclined to think do not correspond to genuine possibilities. Take Norton's dome in Newtonian mechanics for instance (Norton [2003]), or models of GR with closed time-like curves. But is the mere fact that a model is "weird" enough for us to interpret it as unphysical? Motivationalists about physical possibility will say no: they maintain that only a reformulation that eliminates the models in question can license such a judgment.

Chapter 6 gives a motivationalist treatment of theoretical equivalence. It has been argued that theories can only be interpreted as equivalent if they can be shown to share a common

core theory (Read and Moller-Nielsen [2020a]). I develop the details of this common core account of equivalence and defend it against objections.

Finally, Chapter 7 considers whether the inseparability of fact and convention poses a problem for motivationalism. Hilary Putnam [1977, 1980, 1987, 2001] in his work on a phenomenon called conceptual relativity argued that since our theories inevitably involve certain choices of representational convention, one ends up with "equivalent but incompatible" descriptions of the world: theory pairs that by all accounts describe the facts equally well, but which cannot be combined on pain of conceptual confusion or outright inconsistency. Can motivationalism make sense of this phenomenon, or does conceptual relativity show that judgments of equivalence can and must sometimes be made without a coherent account of the reality that grounds the equivalence? In defence of motivationalism, I argue first that motivationalists can concede the ineliminability of representational conventions, second that the common core account of equivalence can handle many alleged cases of conceptual relativity, and third that in recalcitrant cases, the alleged equivalence between the theories involved cannot be known even if it in fact obtains.

2 Preliminaries: What is it to interpret a theory?

1 Introduction

The debate between motivationalists and interpretationalists has the following contours: interpretationalists maintain that it is permissible *ab initio* to interpret symmetry-related models as equivalent. Motivationalists on the other hand maintain that such an interpretation must be underwritten by either a reformulation of the theory that eliminates the symmetry, or at least an account of the shared ontology of the symmetry-related models which grounds their equivalence (Moller-Nielsen [2017]). So, what is at issue in the debate between motivationalists and interpretationalists are the conditions under which it is permissible to interpret symmetry-related models and duality-related theories as equivalent. Sometimes, it is specified that the relevant sense of equivalence here is physical equivalence (see e.g. Martens and Read [2020]). But what exactly is it to "interpret as equivalent"? What is physical equivalence? And, given the importance motivationalists assign to reformulation, what exactly is the role of reformulation in the interpretation of theories or models? These are the questions that we will seek to get a handle on in this chapter, as doing so is a prerequisite for any attempt at settling the debate between motivationalists and interpretationalists.

It is worth dwelling on these questions for some while, since the participants in the debate have not been terribly clear on what mean when they talk about interpreting models as equivalent. At the same time, the notion of interpretation is fraught with difficulties: there are many tropes about interpretation that float around the philosophical literature, and it is not clear that they all fit neatly into a single account of interpretation. For one, there is a common way to think about interpreting a theory according to which interpretation amounts to figuring out what the theory says about the world. In other words, interpretation is something like "theory exegesis": to interpret a theory is to inquire into the theory's content. In philosophical jargon, the task is to determine the truth conditions of the theory (see e.g. van Fraassen [1991], Belot [1998]).

At the same time, we have a ubiquitous distinction between "interpreted theories" and "uninterpreted formalisms". This way of speaking implies that to interpret a theory is to give a formalism content in the first place – say, by putting it into correspondence with

physical target systems. So, already we find a semantic and a metasemantic sense of interpretation. The latter is about putting meaning into the formalism, the former is about getting meaning out. Of these two, the semantic, or hermeneutic sense of interpretation is clearly a better approximation to what motivationalists and interpretationalists argue about. So, for now, we can bracket the metasemantic sense of interpretation, although we will briefly return to it towards the end of the chapter.

There would be little for us to do if the hermeneutic sense of interpretation were clearly the one operational in the debate around interpretationalism. Perhaps we would attempt to give a more detailed and precise description of the hermeneutic account, but nothing beyond that would be required. However, we will see that the semantic way of thinking about interpretation cannot fully make sense of the debate between motivationalists and interpretationalists. One complication arises from the fact that while the hermeneutical account construes interpretation as concerned with the semantics of theories, the disagreement between motivationalists and interpretationalists is clearly at least in part one over what it is rational to believe: to oversimplify somewhat, a motivationalist's initial attitude will be to believe that the structural differences between symmetry-related models of a theory track different ways things could be in the world, whereas an interpretationalist will take a symmetry as proof that some of the models' structure cannot have a physical correlate. But what exactly does this disagreement have to do with our interpretation of the theory's models as equivalent or inequivalent? After all, our attitude towards the relation between the theory and the world is a doxastic matter, whereas the theory's interpretation is putatively a semantic one. At least, the tropes about interpretation we mentioned above suggest that it is. Given how central questions of what it is rational to believe are to the interpretive debates however, one might be tempted to think that theory interpretation is primarily about figuring out what our theories give us reason to believe about the world, and only secondarily about the semantics of our theories. We thus have two broad proposals for how best to think about theory interpretation: one that takes interpretation to be about determining the content of a theory, and a second one that construes interpretation as forming beliefs about the world on the basis of the theory.

The structure of this chapter is as follows: in the next section, I aim to lay the groundworks for the discussion by getting clearer on the notion of physical equivalence. Section 3 will then present the hermeneutic and the doxastic account of interpretation in some detail and

investigate to what extent they help us clarify the dispute between motivationalists and interpretationalists. I will argue that neither one of these standard ways of thinking about interpretation can make satisfactory sense of the debate. So, in order to understand what the debate is about, we will have to go beyond these accounts. That is the aim of Section 4. There, I introduce a third sense of interpretation, revolutionary or ameliorative interpretation, which will help us make sense of the debate and of the role of reformulation within it.

2 Physical Equivalence

What it is for two models to be physically equivalent is for them to agree on all physical structure. But that raises the question: what is physical structure? Relatedly, what is it to interpret structure as physical? I propose that we distinguish two ways in which structure found in a theory can be physical, which I will label the *external* and the *internal* sense.

A piece of structure is externally physical if it has a correlate in the physical world, i.e., if it adequately models some of the structure of the world. For instance, if it turned out that there really is a physical correlate of the vector potential, then its representation in the models of electrodynamics would be externally physical. A piece of structure is internally physical on the other hand if it purports to directly and fully correspond to some structure in the world *according to our interpretation of the theory*; in other words, if our interpretation of the theory has it that the piece of structure in question is meant to serve a direct representational role. On a literal reading of Newtonian Gravitation Theory for instance, the standard of rest is internally physical, since the theory claims that it directly corresponds to a feature of the world. The standard of rest is however not externally physical, since it does not in fact pick out anything in the physical world (at least to the best of our knowledge). Plausibly, not all structure in a theory must be internally physical: for one, many theories contain elements that are uncontroversially gauge. The vector potential of electrodynamics is interpreted by many as not internally physical, but rather as a convenient way to encode information about the electric and magnetic fields. Moreover, according to certain interpretations of quantum mechanics, the wavefunction does not purport to represent any object in the world, but rather encodes information about, say,

which possible states the system could be in given the information we have about it (Maudlin [2019], Ch. 3).

Whether or not a piece of structure is physical in either the internal or the external sense can be the subject of interpretive debates. Take for instance 'dephlogisticated air' as it occurs in phlogiston theory. Uncontroversially, dephlogisticated air is internally physical: according to phlogiston theory, there is such a thing as dephlogisticated air out there in the world. In this case, the difficult interpretive work goes into deciding whether it is externally physical, too. Some have argued that it is and that dephlogisticated air is oxygen (see Schurz [2011]).

When it comes to the wavefunction in quantum mechanics on the other hand, much of the interpretive debate has revolved around the question of internal physicality. On some interpretations, the wavefunction is not itself physical but rather encodes information about which properties are determinately possessed by the system described by the wavefunction: a particle determinately possesses a certain z-spin for instance just in case it is in the corresponding eigenstate. Other views on which the wavefunction is not understood as internally physical are ψ-epistemic interpretations.[6] On the other side of the divide we have the Everett interpretation, Bohmian Mechanics, and dynamical collapse theories.[7] According to these interpretations, the quantum state is internally physical, since it purports to directly represent a feature of the physical world.

In the debate between motivationalists and interpretationalists, the disagreement over the status of symmetry-variant structure often concerns both its internal and external physicality. Some interpretationalists for instance will maintain that symmetry-variant structure can and should be understood as gauge *ab initio*, i.e., even in the absence of a way to neatly isolate the physical structure from the gauge structure. In other words, such interpretationalists would negate the viability of interpreting symmetry-variant structure as internally physical. In support of this claim, they might insist for instance that no coherent interpretation of a theory could countenance violations of Earman's symmetry principles, so that an interpretation of symmetry-variant structure as surplus is forced on us from the

[6] For discussion of ψ-epistemic interpretations of quantum mechanics, see e.g. Leifer [2014] or Maudlin [2019]

[7] See Wallace [2012] for an exposition of the Everett interpretation; Dürr, Goldstein and Zanghi [1996] for Bohmian mechanics, and Ghirardi, Rimini and Weber [1986] for dynamical collapse theories

beginning (Myrvold [2019]). This argument however is in turn grounded in a conviction that the existence, i.e., the external physicality, of dynamically inefficacious spacetime structure is either conceptually incoherent or metaphysically impossible. Motivationalists on the other hand will maintain both that a theory can consistently posit symmetry-variant structure and that it can be rational to believe in the existence of such structure. Thus, Dasgupta for instance states that "it was rational for Newton to believe in absolute velocity even though he knew that it was variant in [NGT] and undetectable. The reason this was rational for him was that he had no good alternative theory to hand. He had good reason (his bucket argument) to think that relationalism was not empirically adequate." (Dasgupta [2011], pp. 853-4)

3 Semantic and doxastic interpretation

Having clarified the notion of physical equivalence, let us now turn towards the two predominant accounts of interpretation, the semantic or hermeneutic account and the doxastic account. We will begin with the semantic construal of interpretation. According to this view, interpretation is theory hermeneutics. To interpret a theory is to determine the content of the theory. This way of thinking about interpretation is ubiquitous in the literature. It is advanced inter alia by van Fraassen and Belot:

> Hence we come to the question of interpretation: Under what conditions is the theory true? What does it say the world is like? These two questions are the same.
>
> (Van Fraassen [1991], p. 242)

> To interpret a theory is to describe the possible worlds about which it is the literal truth. Thus, an interpretation of electromagnetism seems to tell us about a possible world which we know to be distinct from our own.
>
> (Belot [1998], p. 532)

Hermeneutical interpretation is thus the act of explicating the content of a theory; of bringing out clearly what the theory says. One particularly important aspect of this for our purposes is that interpretation of a theory might involve declaring that some of the theory's structure is to be understood as gauge *by the lights of the theory itself*. For a concrete example, take Knox's verdict that "[NGT] itself is best interpreted as a curved spacetime theory, albeit written in a form that obscures its geometrical structure" (Knox [2014], p. 878). Contrary to appearances, Knox tells us, NGT said all along that spacetime was

curved. This kind of interpretive move is perfectly illustrative of Burgess and Rosen's [1997] characterisation of hermeneutic interpretation:

> On what may be called the **hermeneutic** conception, the claim is [...], 'All anyone really means—all the words really mean—is. . . ' (here [...] giving the reconstrual or reinterpretation). Reconstrual or reinterpretation is taken to be an analysis of what really 'deep down' the words of current theories have meant all along, despite appearances 'on the surface'.
>
> (Burgess & Rosen [1997], pp. 6-7, emphasis in original)

Hermeneutical interpretation often takes the form of what Dewar [2023a] labels "external" interpretation. By this he means translation of a theory into a familiar idiom in a way that is faithful to content. The purpose of the translation is of course to elucidate the content of the theory. This already points to a natural place for reformulation in the semantic account of interpretation. If we think about interpretation along hermeneutical lines, we are going to assign a clarificatory role to reformulation: ideally, reformulation recasts the content of the original theory in a more perspicuous light, so that we can understand the theory more readily. What is crucial here is that reformulation should preserve semantic content.

We can try to understand what motivationalists and interpretationalists mean by "interpreting models as equivalent" along hermeneutic lines. According to this construal, when we interpret two models as equivalent, we judge that they have the same semantic content. In practice, this might amount to believing that any symmetry-variant structure is gauge by the lights of the theory that exhibits it.[8] One advantage this construal has is that it is nicely aligned with the tropes about interpretation we presented above. In other words, it sticks closely to the predominant way of thinking about interpretation. Moreover, while not primarily concerned with our doxastic attitudes, it seems to be able to accommodate the intuition that theory interpretation plays a significant role in our formation of beliefs about the world. It can do so because a bridge between the semantic and the doxastic is provided by a background commitment to scientific realism: realism says we ought to believe that our best scientific theories are approximately true. In other words, the content of our beliefs ought to coincide with the content of our best theories. If we stick to this semantic way of thinking about theory interpretation, we could therefore understand the motivationalist and the interpretationalist to be advancing different hypotheses regarding the semantic content of our scientific theories. Oversimplifying somewhat, the interpretationalist takes himself to be at liberty to view any symmetry-related models as

[8] Dewar and Eva [unpublished] can be understood as proposing something of this kind.

having the same content, whereas the motivationalist believes that for such a judgment to be warranted, a paraphrastic reformulation of the theory is needed. On this construal, the disagreement over appropriate doxastic attitudes would be downstream of a semantic dispute, and would arise modulo a shared commitment to scientific realism.

But there are some worries about this way of understanding the debate over when models may legitimately be interpreted as equivalent: first, it drags us into very murky, obscure questions about the semantic content of classes of models. For instance, one will be steered towards trying to decide whether, say, Newtonian Gravitation Theory has the same semantic content as Newton-Cartan Theory and so forth. While some have attempted to tackle this question (Glymour [1977], Knox [2014], Weatherall [2016b], March [2024]) and relevantly similar ones, they seem intractable, and it is difficult to see how we could adjudicate them. Moreover, to the extent that one answer is more plausible than the other, it is that they are semantically *in*equivalent. As Butterfield [2018] emphasizes, Newton clearly believed in the full structure of absolute space and time (and for good reasons, too – recall the bucket argument). Unless we are prepared to say that he was severely mistaken about the content of his own theory, Newtonian Gravitation Theory, at least as Newton understood it, is semantically inequivalent to Newton-Cartan Theory. But if this is right, then the hermeneutic approach cannot capture a sense in which Newton-Cartan Theory, or perhaps Maxwell Gravitation, is the *best* interpretation of Newtonian Gravitation Theory. I maintain however that we want an approach to interpretation that does allow us to say this. The approach developed below does so. A final concern is this: the hermeneutic approach does not place the disagreement over what it is rational to believe about the world in light of our scientific theories at the centre of the debate between interpretationalists and motivationalists, even though that is by far the most important and interesting point of contention.

An alternative way of thinking about interpretation, the doxastic way, arguably does better. According to this account, theory interpretation is broadly about deciding what to believe about how things are in the world in light of relevant features of the theory to be interpreted. These features can include empirical success and limitations, pathological features such as inconsistencies or singularities, a mismatch between spacetime and dynamical symmetries, and much more. I take Redhead to be advancing a broadly doxastic approach to interpretation when he writes that

> [An interpretation of QM] is simply some account of the nature of the external world and/or our epistemological relation to it that serves to *explain* how it is that the statistical regularities predicted by the formalism with the minimal statistical interpretation come out the way they do.
>
> (Redhead [1989], p. 44)

Of course, an interpretation of a theory in the doxastic sense can take the semantic content of the theory into account. In the very simplest case, in which one has a theory with clearly delineated semantic content and unlimited empirical success, doxastic interpretation would presumably amount to simply believing the theory at face value. An advantage of the doxastic approach however is that it does not limit itself to semantic considerations. Part of our interpretation of quantum mechanics for instance might be that the world has an emergent branching structure (Wallace [2012]). But we don't need to ponder whether that is part of the "literal semantic content" of quantum mechanics – indeed, that would seem like a fruitless endeavour. According to the doxastic way of thinking about interpretation, to interpret two models as equivalent simply is to judge that at most the structure common to the models is physical. Some advantages of the doxastic account are the following: this construal of what it is to interpret models as equivalent immediately gives centre stage to the dispute between motivationalists and interpretationalists over what it is rational to believe about the world. It also allows us by and large to steer clear of tiresome debates over the semantics of scientific theories. In these respects, it is an improvement over the hermeneutic account.

However, it faces certain problems of its own. The most important of these is that it cannot make sense of what we are up to when we interpret theories that have been superseded. Take for instance the debate over how to interpret Newtonian Gravitation Theory. Part of what is disputed here is whether Maxwell Gravitation or Newton-Cartan Theory provides the best reformulation of NGT. This is taken by many to be an interpretive dispute. But clearly, it is not a dispute over what structure to regard as physical: NGT has long been superseded, and plausibly, neither MG nor NCT posit any structure that we ought to take to be physical. When it comes to the interpretation of obsolete theories, the first account we gave of interpretation, according to which interpretive judgments are judgments of semantic content and equivalence, fares much better than the doxastic approach. Secondly, there is the problem that there does not seem to be an entirely natural place for reformulation in the doxastic account of interpretation. To the extent that it does play a role, its function would presumably be to codify at least some of what one has concluded

during the interpretation of the theory. But of course, normally, what one ought to believe about the world on the basis of a theory's successes and failures will neither be codifiable in a set of models, nor will it coincide with the content of what we usually regard as appropriate reformulations of our theories. For instance, while Newton-Cartan Theory is widely seen as a reformulation of Newtonian Gravitation Theory, it is not what one ought to believe about the world given the empirical success of NGT within its domain of applicability.

The situation we find ourselves in therefore is the following: before we can even begin to settle the dispute between motivationalists and interpretationalists, we must get a handle on the notion of interpreting models as equivalent. However, neither of the two standard construals of this notion, i.e., neither the semantic construal nor the doxastic construal, is free of problems. The semantic conception struggles to accommodate interpretationalists who, in regarding a theory's symmetry-related models as equivalent, are not concerned with whether the models in question really have the same semantic content, but with whether they agree on all physical structure. The doxastic account of interpretation on the other hand seemingly cannot make sense of the debate between motivationalists and interpretationalists as directed towards the interpretation of obsolete theories. This matters because motivationalists and interpretationalists will disagree strongly about the correct attitude to take towards, say, Newtonian Gravitation Theory in the absence of reformulations such as Newton-Cartan Theory or Maxwell Gravitation, even though they might agree that whatever the correct interpretation of the theory is, it is unlikely to describe any physical structure in the world. Here, a semantic construal of the debate makes more sense: motivationalists and interpretationalists can be taken to differ over the content of the theory, even though they agree that whatever the content, the theory is false.

Since neither account of interpretation is fully satisfactory, it seems then that we are already stuck at the stage of clarifying the dispute. But progress can be made. I will introduce a different way of thinking about interpretation, called ameliorative or revolutionary interpretation, which I will argue allows us to make sense of the debate around interpretationalism. We will turn to this in section 4. But first, we will try to get clearer on the idea of judging two models to be physically equivalent. We do this both to flesh out the details of the doxastic account and also because it is a central notion in the

debate between interpretationalists and motivationalists and requires elucidation for that reason.

4 Revolutionary interpretation

We have attempted to clarify some of the terms of the debate between motivationalists and interpretationalists. What has eluded us so far however is a way to think about theory interpretation that fully makes sense of the debate. As we saw, neither the hermeneutic nor the doxastic account seems to adequately capture what the participants in the debate mean by interpretation. This is not to say of course that there is anything wrong with hermeneutic or epistemic interpretation. Each raises a set of important, worthwhile questions about theories. The claim is merely that neither kind of interpretation seems to match the one operational in the debate between motivationalists and interpretationalists. Therefore, I want to put a new sense of interpretation on the table: revolutionary,[9] or ameliorative interpretation.

Revolutionary, or ameliorative interpretation is about finding a superior alternative to an existing theory beset with conceptual problems. In short, it is about fixing a theory's flaws – hence the label 'ameliorative'. In more detail, interpretation here amounts to treating an analysis of a theory's flaws – a mismatch between dynamical and spacetime theories, for

[9] This terminology is taken from Burgess and Rosen's work on nominalism (see Burgess & Rosen [1997]). Burgess and Rosen draw a distinction between hermeneutic and revolutionary nominalism: revolutionary nominalists believe that mathematical talk is committed to abstract objects and therefore want to replace it with a superior idiom that eschews such commitment. Hermeneutic nominalists on the other hand want to show that our usual mathematical parlance does not commit us to abstract objects in the first place. They might argue for instance that whenever I say 'The number of planets is two', all I really mean is 'There are two planets', and so forth. If so, there is no need to revise our language and we can go on speaking of numbers, vector spaces, manifolds etc. just as before. Clearly, Burgess and Rosen's distinction maps onto the debate between motivationalists and interpretationalists at least to some extent: one might have a motivationalist for instance who believes that NGT is committed to a standard of rest and therefore wants to reformulate the theory so as to eliminate this commitment. Such a motivationalist is thus similar in attitude to the revolutionary nominalist. An interpretationalist however might want to say that NGT isn't committed to a standard of rest in the first place, despite appearances. Once NGT is understood properly, it will be seen that the standard of rest is gauge. This kind of interpretationalist resembles the hermeneutic nominalist. A word of caution about this analogy however: while some motivationalists and interpretationalists may think about their attitudes in this way, by no means all do. For instance, some interpretationalists might grant that NGT really does posit a standard of rest, but might think that the symmetry-variance of the standard of rest immediately demonstrates that it cannot be physical.

instance – as a hint at an improved version of the theory. To interpret a theory in this sense is to discern an alternative hidden within its structure. As a project, revolutionary interpretation is unencumbered by any requirements to yield a theory equivalent to the original. The new theory certainly need not be semantically equivalent to the old theory – indeed, if it were, it would be difficult to see how amelioration could be said to have taken place. But as I will argue in Chapter 4, revolutionary interpretation can go so far as to yield theories that are empirically inequivalent to the theory we treat as our starting point. What matters is that the new theory is superior to the old.

An example of this at work is perhaps furnished by "interpretations" of quantum mechanics that modify or amend the formalism to correct an alleged incompleteness – take Bohmian Mechanics or dynamical collapse theories for instance. A very clear example of revolutionary interpretation is the replacement of a force-based formulation of NGT with a potential-based formulation for the purposes of modelling constant positive mass densities throughout the universe (see Malament [1995]). In this example, which we will discuss in greater detail in Ch. 5, the theory's mathematical framework is revised so as to allow us to model systems the original theory could not adequately capture. A further example of revolutionary interpretation, from mathematics rather than physics, is the Cauchy-Weierstrass reformulation of the theory of infinitesimals, which fixed the latter theory's inconsistencies and conceptual issues.

Crucially, I believe that reduction and (internal) sophistication[10] are best thought of as forms of revolutionary interpretation: a theory exhibiting surplus structure is replaced by an alternative, inequivalent theory that is free of that structure, but preserves the symmetry-invariant features of the original theory.[11] As van Fraassen and Ismael state, "[t]here are two stages in theory construction. The first is to generate a set of models rich enough to embed the phenomena, the second is to attempt to simplify those models by exposing and eliminating excess structure. Continuing in this way the structure of the models is pared down, being careful not to jeopardize their capacity to embed the phenomena" (Ismael and van Frassen [2002], p. 390).

[10] Whether so-called *external sophistication* (see Dewar [2019], Martens and Read [2020]) should be regarded as a form of revolutionary interpretation is a trickier question. I address it later in this chapter as well as in Chapter 3, and argue that it is unlikely external sophistication can be made to meet the requirements one is going to have for revolutionary interpretation.
[11] For discussion of reduction and sophistication, see Dewar [2019].

As such, the move from NGT to NCT for instance is also to be understood as a case of revolutionary interpretation. Indeed, Knox's [2014] guiding question, "What is the ideal spacetime setting for Newtonian Gravitation Theory?", is precisely the kind of question a revolutionary interpreter would ask. In contrast to the hermeneutic and doxastic approaches, the revolutionary's question aims not at semantic content or lessons about the nature of reality, but instead at optimisation of the theory.

What I want to suggest therefore is that when interpretationalists and motivationalists talk about theory interpretation, they have in mind, at least implicitly, something akin to revolutionary interpretation. Or if they don't, then they should: the clearest and most fruitful way to construe their disagreement is as about revolutionary interpretation. For an instructive example, take Dewar's [2023] preferred approach to interpretation, which he calls *internal interpretation*. As I will argue, internal interpretation can be understood along revolutionary lines. This approach to interpretation is especially relevant to the debate around motivationalism, since the aim of internal interpretation as Dewar conceives of it is to find the best way of divvying up a theory's structure into physical and artefactual. This amounts to deciding which models of a theory are to be regarded as equivalent, of course, and so it is precisely such judgments of (in-)equivalence for models that according to Dewar constitute an internal interpretation. The details of Dewar's proposal are congenial to the interpretationalist – at its heart is the idea that one can legitimately interpret a theory by simply declaring models equivalent on the basis that they are related by a symmetry. Nonetheless, his broad ideas that interpretation is about identifying grounds on which models may be regarded as equivalent, and that an interpretation is to be judged by whether the (in-)equivalences it declares to hold between models are plausible, are ones that unite interpretationalists and motivationalists.

Now, on Dewar's proposal of internal interpretation we can always trivially interpret symmetry-related models as equivalent since there is supposed to be a cheap, universally available, fail-safe method which allows us to always construct such an interpretation. This method is called *external sophistication*.[12] To externally sophisticate a theory is to declare the symmetry-related models of the theory isomorphic or indeed equivalent by fiat, even if they fail to be isomorphic *qua* mathematical objects (see Dewar [2019], [2023], Martens and Read [2020]). Often, one attempts to make this idea more precise using

[12] The terminology here – internal interpretation via external sophistication – invites some confusion, of course, but is correct as stated.

category theoretic tools (see Dewar [2023], as well as Ch. 3 of this thesis). External sophistication contrasts with internal sophistication, which amounts to reformulating a theory in a mathematically rigorous way so that the symmetry-related models are transformed into (set-theoretically) isomorphic models (see Martens and Read [2020]). This idea of external sophistication now allows us to reconstrue the debate between motivationalists and interpretationalists along "revolutionary" lines.

Recast explicitly in terms of revolutionary interpretation, the disagreement between the two camps comes down to how straightforward it is to give a revolutionary interpretation of a theory that eliminates the theory's symmetries.[13] Barring some nuance (see Luc [2023]), interpretationalists are confident that for any theory that exhibits symmetry-variant structure, there is a superior theory out there that gives us such an interpretation. External sophistication is one way to construct such a theory. Why this is contentious however is that whereas internal sophistication is a paradigm of revolutionary interpretation, external sophistication is a much more dubious method. Motivationalists therefore will insist that the existence of a reformulation that eliminates a theory's symmetries is not guaranteed and must be explicitly demonstrated via a method other than external sophistication. Only once that is done, and a successful ameliorative interpretation has explicitly been given via more rigorous methods such as internal sophistication, are we at liberty to regard the original theory's account of the world as superseded.

The issue therefore becomes whether an appeal to external sophistication can vindicate interpretationalism. As I have conceived of it here, this boils down to asking whether external sophistication is a legitimate way to give a revolutionary interpretation of a theory. In Chapter 3, I will argue that it is not. To be legitimate, external sophistication must deliver a number of things. Minimally, it must yield a new theory with determinate semantic content. Martens and Read [2020] have argued, convincingly to my mind, that this is doubtful. In Chapter 3, I will give further reasons for thinking that in general, external sophistication fails to fix a unique, determinate content to be ascribed to the theory putatively being interpreted. But even when it does, or even if the method could be refined in such a way as to always uniquely define a content for the externally sophisticated theory, there remains a second problem. For external sophistication to constitute successful

[13] To be clear, what follows is not meant to necessarily reflect the actual beliefs of interpretationalists and motivationalists. Rather, it is meant as a rational reconstruction of the debate.

ameliorative interpretation, it must be that the sophisticated theory really is a better theory than its predecessor, as measured by the standard criteria for theory evaluation such as explanatory and unificatory power, elegance, simplicity, etc. The issue, I will argue, is that even assuming external sophistication determines a well-defined content for the resulting theory, that content will remain completely opaque. This in turn means that it will be difficult, if not impossible, to ascertain whether the sophisticated theory really improves on its unsophisticated predecessor. In consequence, we will not have a warrant for believing the new theory over the old. Overall then, external sophistication can (and arguably does) fail as a vindication of interpretationalism on at least two levels. On a semantic level, it is unclear whether external sophistication can secure a determinate content for the sophisticated theory. On an epistemic level, we have no way of knowing whether the sophisticated theory is superior to the original theory, and therefore no warrant for dismissing the latter.

This sketch of the debate between interpretationalists and motivationalists as seen through the lens of revolutionary interpretation oversimplifies of course. In particular, it represents only one strand of interpretationalism. One group of interpretationalists it neglects is those Sider [2020] calls quotienters: far from being confident that reformulation must invariably be possible, quotienters think that even when two models genuinely are equivalent, the demand for reformulation potentially cannot be met and is therefore unreasonable.[14] Quotienters urge us to mind the ubiquity of and importance of artifacts of representation, which they think renders revolutionary interpretation, understood here as the project of eliminating these artifacts, an elusive, frequently unattainable goal. So, while some are interpretationalists because they consider the availability of reformulations to be a foregone conclusion, quotienters are interpretationalists because they think that even in cases in which it is blindingly obvious that two models are equivalent, certain potentially ineliminable artifacts of interpretation may render them non-isomorphic. We will defer consideration of the quotienting view until Chapter 7. But for now, it should be noted that the quotienter's view can also be fruitfully construed as one that concerns revolutionary interpretation: quotienters think that revolutionary interpretation is in many cases bound to fail because there will often be indispensable but representationally idle artifacts of representation.

[14] Note however that a notion of "cheap" reformulation, such as external sophistication, may be quite congenial to the quotienter's position.

Furthermore, note that besides giving us a clear construal of the debate between motivationalists and interpretationalists, the concept of revolutionary interpretation sidesteps many of the issues that beset hermeneutic and doxastic interpretation. First, a role for reformulation is immediately built into the account. Secondly, in contrast to the hermeneutic account, this role is not confined to the rather opaque task of restating the theory's semantic content. The role of reformulation in revolutionary interpretation is not to present the content of a theory in a faithful yet more perspicuous way, but rather to provide a new theory that is suitably continuous with the old one but improves on it in various ways. In this respect, revolutionary interpretation resembles doxastic interpretation more than hermeneutic interpretation. We will see this in much detail in Chapter 4, where I present a reduction of Newtonian Gravitation Theory with respect to its scale symmetry. The resulting theory is inequivalent to NGT by any reasonable standard, so that the project I am engaged in there is most fruitfully conceived of as an instance of revolutionary interpretation, not hermeneutic interpretation.

Finally, another advantage of the revolutionary account is that it is applicable both to our current best theories and to obsolete ones. After all, even when it comes to obsolete theories, there is a clear sense in which they can be improved by their own internal standards: the mismatch between dynamical and spacetime symmetries in NGT for instance is a blemish *by the lights of the theory itself*, and so can be the target of ameliorative interpretation. So, the revolutionary account of interpretation overcomes all of the serious problems its competitors are facing when it comes to making sense of the debate between interpretationalists and motivationalists.

That being said, as mentioned above, revolutionary interpretation is not meant as a replacement for hermeneutic or doxastic interpretation, both of which are important philosophical activities. Rather, we should think of revolutionary interpretation as complementary to these other forms. It seems to me that they can interact fruitfully with one another. For one, hermeneutic and doxastic interpretation can help us identify and diagnose the flaws in a theory that we then seek to remedy through revolutionary interpretation. As such, we can even imagine alternating patterns of interpretation: hermeneutic followed by revolutionary, with the result of a new theory that in turn is in need of hermeneutic interpretation, and so forth.

This largely concludes my discussion of how best to understand the notion of theory interpretation that underlies the debate between interpretationalists and motivationalists. One part of the motivationalist doctrine still requires clarification however: namely, motivationalists maintain that we are permitted to interpret symmetry-related models as equivalent once we have found a reformulation that eliminates the structural differences. In those cases, we are licensed to interpret the variant structure as mere gauge.

The best way, I think, to make sense of what is going on in these cases of "reinterpretation" invokes the aforementioned third sense of interpretation, which is metasemantic, or meaning-determining. This is the sense operational in the distinction between interpreted and uninterpreted formalisms. What is meant by interpretation here is giving a formalism content by putting it in correspondence, directly or mediately, with physical systems. This is what I think we do when we interpret a theory with symmetries in light of its reformulation: we stipulate that the formalism of the old theory is now to be understood as merely a representation of the reformulated theory. This gives us a new interpretation of the formalism, in addition to the original, "literal" interpretation. This is what Coffey has in mind, I take it, when he discusses the different interpretations of NGT available to us:

> If, following Earman ([1993]), we take seriously the flat geometric space-time structure of Newtonian gravitation theory and its posited force field, and also the dynamical non-flat space-time structure of Cartan theory, then the theories they represent are distinct, for the fundamental ontologies they offer are different. If, following Barrett ([2008]), we do not take the background mathematical structure of Newtonian gravitation theory as representing something primitive, then we're free to understand that formulation as an overwrought description of Cartan theory (suitably interpreted). But it's not as though Earman and Barrett use different concepts of theoretical equivalence. Rather, they are choosing to interpret the formalisms differently.
>
> (Coffey [2014], p. 835)

In other words, we are at liberty to stipulate the interpretation of the formalism of NGT: we can interpret it literally or as an "overwrought" representation of NCT. Motivationalists emphasize however, in disagreement with interpretationalists, that this liberty is granted to us only through a demonstration that NCT is a reformulated version of NGT.

Having clarified how we ought to think about theory interpretation in the context of the dispute between interpretationalists and motivationalists, we can now turn to examining this dispute in detail. This will be our task for the next chapter.

3 Motivationalism: an examination

1 Introduction

My undertaking in this chapter is a thorough examination of motivationalism. I will develop what I consider to be a motivationalism more plausible and robust than extant ones. Section 2 presents the definition of symmetries we will be presupposing. Section 3 gives a detailed overview of motivationalism, focussing particularly on choice points that differentiate versions of this doctrine. Section 4 defends motivationalism against opposed views such as interpretationalism. Special attention is paid to a recent interpretationalist strategy for theory interpretation called external sophistication. I argue that external sophistication is not a viable method of interpretation. Section 5 stakes out my own position within the motivationalist camp. My disagreement with other motivationalists is over the conditions under which it is permissible to interpret symmetry-related models as equivalent. Relative to other versions of motivationalism, I emphasize the need for reformulation, but deemphasize the need for a perspicuous metaphysical picture of the ontology common to symmetry-related models. I also argue against placing certain very demanding conditions on reformulation, most importantly the condition of intrinsicality defended by Jacobs [2022a].

In this chapter, I am exclusively concerned with motivationalism about symmetries. As mentioned in in the introduction to this thesis, one can however also be a motivationalist about dual theories. On (one version of) this view, dual theories may only be regarded as theoretically equivalent if we have an account of the ontology they share. I defer discussing motivationalism about dual theories until Chapter 6 of this thesis. In particular, when discussing objections to motivationalism in the current chapter, I will not mention the distinctive problems motivationalism about dual theories faces.

2 Symmetries

As it turns out, it remains controversial how best to understand the notion of a symmetry. A symmetry is a transformation of a theory's models that is *closed* in the sense that it

transforms any model of the theory into a structure that is also a model of the theory (see Earman [1989]). In other words, a symmetry transformation never takes us outside the space of the theory's models. As a definition of symmetry however this will not suffice as it includes too many transformations by far (Belot [2013]). For a useful, informative definition of symmetry, further conditions will have to be imposed. This is where the controversy lies. Common requirements include that a symmetry transformation should preserve some (or even all) physical quantities (e.g. Read and Moller-Nielsen [2020b]) and that it should map every model to an empirically equivalent one (e.g. Dasgupta [2016]). None of these conditions is universally accepted however.

Whatever definition of symmetry we settle on is therefore bound to invite objections. There is no way to fully extricate ourselves from the debate. Nonetheless, I do not intend to delve into it very deeply. I certainly do not propose to offer a full, definitive resolution of the controversy. Indeed, I will not even spell out my preferred way of understanding symmetries at a level of detail that would cover every point of contention. Luckily, the disagreement between interpretationalists and motivationalists is not sensitive to every such choice point, so that there is nothing problematic about proceeding in this way. Some proposed definitions of symmetry must be rejected, to be sure; specifically those that have built in the requirement that any two symmetry-related models be physically equivalent.[15] This would render the interpretationalist position correct by definition. Rejecting this criterion however does not narrow down the field of proposals by much. Let me therefore put my preferred way of thinking about theories and symmetries on the table and then briefly situate it within the debate over the definition of symmetries.

Throughout this thesis, I will understand theories and symmetries according to the following schema: A theory is a class of models. By a model I mean a tuple of sets, each potentially equipped with geometrical structure, and mappings between these sets. One commonly distinguishes between the kinematically possible models (KPMs) of a theory, which are all the models containing the right kinds of sets and functions, and the dynamically possible models (DPMs), which are the theory proper and are singled out from the KPMs via what are called the theory's *dynamical equations*.

[15] Such a requirement is discussed but rejected by Belot [2013]

Let T be a theory, with models of the form $\langle \mathcal{V}, \phi \rangle$, where $\mathcal{V}$ contains T's structured sets, or value spaces, and ϕ contains the mappings between these sets included in T. Then we say that a dynamical symmetry of T is a mapping from KPMs to KPMs that is (1) induced by a transformation acting on the theory's value spaces and (2) closed on the space of DPMs. In other words, a symmetry is given by an active transformation ψ acting on $\mathcal{V}$, such that the map

$$h_\psi : \langle \mathcal{V}, \phi \rangle \rightarrow \langle \mathcal{V}, \psi \circ \phi \rangle \tag{3.1}$$

maps the space of DPMs onto itself.[16] Here, we call h_ψ the *lift* of the transformation ψ.

Where does this definition place us within the debate over symmetries? Dasgupta [2016] distinguishes three broad approaches to symmetry: the formal, the ontic, and the epistemic. At a first pass, formal definitions of symmetries are those that invoke exclusively mathematical features of the theory's formalism, i.e. of the theory's class of models *qua* uninterpreted, purely mathematical objects. Ontic and epistemic definitions on the other hand involve the interpretation of the theory, specifically the physical features of the states of affairs represented by the theory's models. Thus, on an ontic understanding, a symmetry is a transformation of the theory's models such that certain physical features are preserved across the states of affairs represented by symmetry-related models.[17] According to Dasgupta, these physical features will include the nomic facts, i.e. the facts as to what laws obtain in these states of affairs, and certain privileged physical structure and quantities. Epistemic definitions of symmetries are also couched in terms of the preservation of certain features across states of affairs represented by SRMs. The features required to be preserved are here precisely those determinable in principle by empirical means.[18] Unlike with ontic definitions, these need not be limited to features explicitly represented by the theory (think: velocity, charge, field strength, etc.). They will include any and all measurable or perceivable attributes. On the epistemic approach to symmetries, at most those transformations that exclusively relate empirically equivalent models qualify as symmetries.

Now, the debate between interpretationalists and motivationalists concerns the validity of what are called symmetry-to-reality inferences. The basic pattern of such inferences is:

[16] With slight modifications, this definition is taken from Jacobs [unpublished].
[17] The ontic approach is advocated inter alia by Belot [2013], Read and Moller-Nielsen [2020b], and Wallace [2022]
[18] The epistemic approach to symmetries is pursued inter alia by Dasgupta [2016], Ismael and van Fraassen [2003], and Caulton [2015]

'Models M_1 and M_2 are related by a symmetry. They are therefore equivalent' or 'Quantity Q varies across symmetry-related models. Q is therefore not physical'. The intuition behind these inferences is that any degrees of freedom variation of which never makes a dynamical difference cannot but indicate redundancy. Interpretationalists liberally reason according to these inferences whereas motivationalists deny that they are admissible without further input. But evidently, for them to be plausible at all, they must concern symmetries that only relate empirically equivalent models. This points to the epistemic approach to symmetries as the one most congenial for our purposes. The definition of symmetries I have given however is a purely formal one.[19] The reason for this is that, as Read and Moller-Nielsen [2020b] and Wallace [2022b] point out, the epistemic approach deviates from how symmetries are normally understood in physics, where accounts along formal or ontic lines are predominant. Dasgupta argues in favour of the epistemic approach because he wants to uphold symmetry-to-reality inferences in unqualified form. Methodologically, it strikes me as preferable however to opt for the pristine, purely formal definition of symmetries presented above and to restrict the intended scope of symmetry-to-reality inferences to those symmetries that we take to preserve empirical content.[20]

Thus, the interpretationalist's claim that symmetry-related models are to be invariably and immediately interpreted as equivalent is only supposed to apply to those symmetries that we have determined do relate empirically equivalent models. I will take for granted that all of the particular symmetries discussed in this thesis fall in this category. As far as I can tell, this is in no case controversial – all of the examples to come are widely recognized as transformations turning any model into an observationally indistinguishable one. Therefore, I will leave the assumption that the symmetries under discussion preserve empirical content tacit throughout the thesis.

Finally, a clarification: While I hope that the schema for understanding theories and symmetries just presented is sufficiently general and comprehensive to cover most of the salient examples of symmetries, I make no claim that it is exhaustive in this sense. In light of Belot's [2013] observation that it is very difficult to specify criteria satisfied by all and

[19] Dasgupta [2016] classifies definitions of the kind I have given as ontic due to the fact that they require preservation of the laws across symmetry transformations. But this requirement can be captured in purely formal terms, by demanding that solutions of the dynamical equations never be mapped to non-solutions. So, *pace* Dasgupta, I believe that my definition of symmetries is an instance of the formal approach.

[20] The discussion in Read and Moller-Nielsen [2020b] lends much support to this way of proceeding.

only the transformations commonly considered symmetries, Read and Moller-Nielsen [2020b] go in for a pluralistic understanding of symmetries. The account offered here is supposed to be consistent with such a pluralism.

3 What is motivationalism, exactly?

Motivationalism addresses the question: under what circumstances is it legitimate to interpret symmetry-related models as equivalent? As such, it is an epistemological thesis – it is concerned with the justification of our beliefs in the equivalence of models. Motivationalism is perhaps best understood as the rejection of its rival, interpretationalism (see Moller-Nielsen [2017]). Interpretationalists maintain that the mere fact that two models are symmetry-related suffices to legitimize the judgment that they are equivalent. Motivationalists argue that this is not enough. While symmetry-relations between models motivate us to interpret these models as equivalent, we need to supply the grounds for this equivalence before we are licenced to do so (Moller-Nielsen [2017], Martens and Read [2020], Read and Moller-Nielsen [2020a, b], Jacobs [2022a]).

In its canonical form, motivationalism involves the following claims (Moller-Nielsen [2017], Martens and Read [2020], Read and Moller-Nielsen [2020a, b], Jacobs [2022a]):

1. If models are symmetry-related, this motivates interpreting them as equivalent
2. To be licenced to interpret symmetry-related models as equivalent, we need "a metaphysically perspicuous characterization of the reality that is alleged to underlie symmetry-related models" (Moller-Nielsen [2017], p. 1256); a "coherent metaphysical picture of the common ontology underpinning their equivalence" (Martens and Read [2020], p. 320)
3. Providing such a picture involves showing that the symmetry-related models either are isomorphic or can be rendered isomorphic (or even identical) through reformulation of the theory
4. The isomorphic models must be presented intrinsically, which is to "lay down a set of relations, functions and operators which explicitly represent the world's physical structure" (Jacobs [2022a], p. 15)

5. There are no defeating "explanatory/metaphysical considerations [that] preclude us from regarding those models as being physically equivalent" (Martens and Read [2020, p. 321]) (see also Martens [2018])

However, not all motivationalists accept every aspect of the canonical view. While motivationalists of all stripes agree that symmetries (a) do not by themselves licence judgments of equivalence, but (b) are a *prima facie* blemish of a theory that motivates their elimination, they disagree among themselves on many points of detail. For instance, not all motivationalists insist on Condition 5, the 'no defeaters' condition.[21] Moreover, some motivationalists accept that we need to perspicuously characterize the ontology common to symmetry-related models before we are permitted to interpret them as equivalent but don't think that this necessarily requires reformulation of the theory – it might suffice for instance to define the quantities invariant under the symmetries. Such a view has been tentatively defended by Read (2022). Yet other disagreements can arise over Jacobs' intrinsicality condition. While authors such as Read, Martens, and Moller-Nielsen insist on a *perspicuous* picture, they have not spelled out their notion of perspicuity in terms of intrinsicality. It is therefore not clear whether they would agree with Jacobs that perspicuity requires intrinsicality.

As this illustrates, motivationalists face many choice points. The principal ones I will focus on in this chapter, and my answers to them, are:

1. Do we need to reformulate in a way that eliminates any symmetries relating non-isomorphic models? Yes
2. Do we need intrinsic reformulation? No
3. Do we need a perspicuous metaphysical picture of the ontology shared by symmetry-related models before we may interpret them as equivalent? Not invariably

My insistence on (1) puts me in disagreement with the view articulated by Read (2022). My answer to (2) brings me into conflict with Jacobs. Finally, my views on (3) are rejected by perhaps all other motivationalists. Nonetheless, I will argue that the version of motivationalism I am defending is more plausible than its rivals. Its perhaps most

[21] See Martens [2018] for discussion

significant difference from other forms of motivationalism concerns question 3: I will argue that while a perspicuous metaphysical picture of a theory's commitments is an ideal we should strive to obtain, there can be cases in which judgments of equivalence for symmetry-related models can be underwritten by a reformulated theory that is not fully perspicuous (or indeed not at all perspicuous), as long as the reformulation meets a condition I dub *'recognisable adequacy'*. I will elaborate on this in much greater detail below. As for Martens' 'no defeaters' condition, I think that it is plainly correct. Many reformulations are simply worse than the theory they replace. For instance, holonomy-based electrodynamics (Healey [2007]) is worse than vector potential electrodynamics, for reasons I will elaborate on later (but which are in any case widely known – see e.g. Jacobs [2024]).

The issues I will focus on in this chapter do not cover every aspect of motivationalism one might disagree over. One further issue on which motivationalists who insist on reformulation are divided is that of which symmetry-related models demand reformulation before they can be considered equivalent – whether it is any distinct ones, or merely any non-isomorphic ones. One popular view is that anti-haecceitism allows us to give a coherent account of the ontology shared by isomorphic models (Moller-Nielsen [2017], Read and Moller-Nielsen [2020b]). But not all motivationalists accept anti-haecceitism (see Dasgupta [2011]). Related to this of course is the question of what kind of reformulation symmetries motivate us to give – whether it is sophistication or reduction. I will not discuss this issue in much detail here. As I will argue in the next chapter, we should look for both reductions and sophistications and choose whichever is best. I reject anti-haecceitism, but I don't think that that necessarily makes sophistications worse than reductions.

In the next section, I will defend motivationalism against other views, such as interpretationalism. Subsequently, I will elaborate on my own preferred version of motivationalism and explain I believe it to be more plausible than other versions.

4 In defence of motivationalism

Motivationalists are trying to carve out a somewhat uneasy status for symmetry-variant structure: on one hand, a blemish of a theory that motivates elimination; on the other hand,

not so bad as to immediately convict a theory of falsity. Our task is to show that this defines a stable position to occupy. Anti-motivationalists maintain that symmetry-related models ought to be interpreted as equivalent by default. But not all anti-motivationalists agree on the reasons for this. In what follows, we will look at three anti-motivationalist views, roughly graded by strength. The first, and strongest, position is Myrvold's [2019]. Against the background of the dynamical approach to spacetime structure, he argues that Earman's well-known symmetry principles SP1 and SP2 have the status of analytic truths, so that it would be incoherent to deny the equivalence of symmetry-related models. A more moderate view is defended by Luc [2023], who maintains that an interpretation of symmetry-related models as equivalent should be the default, but that such a verdict is in principle defeasible. We will discuss these views in Sections 3.1 and 3.2, respectively. Finally, in Section 3.3 we consider an argument to the effect that theories can always immediately be reformulated in a way that renders symmetry-related models isomorphic, using a method called "external sophistication".[22] According to this view, the motivationalist's demand for such a reformulation is not so much superfluous as it is trivial to meet.

4.1 Interpretationalism and the dynamical approach to spacetime structure

The first set of challenges to motivationalism arises from the so-called dynamical approach to spacetime structure, according to which spacetime structure is in a sense a mere "shadow of the dynamics".[23] [24] Slightly less enigmatically, the idea behind the dynamical approach is that attributions of structure to spacetime are abstractions from the dynamics of moving bodies. Thus, to say that spacetime has a certain structure is a way of codifying the transformation properties of the dynamics. This suggests that surplus structure could not possibly be physical, and so puts pressure on a default attitude of discrimination between symmetry-related models. From this point of view, Earman's symmetry principles

[22] External sophistication was first proposed by Dewar (see Dewar [2019])

[23] For the dynamical approach to spacetime structure, see inter alia Brown [2005] and Brown and Read [2021]

[24] While I had already realized that Myrvold's view poses a challenge to motivationalism, the full extent of the tension between motivationalism and the dynamical approach was made vivid to me by Eleanor March. I thank her for it. Note that Read [2022] is aware of the tension and rejects the idea that SP violations are incoherent

(SP1) Any dynamical symmetry of T is a space-time symmetry of T

(SP2) Any space-time symmetry of T is a dynamical symmetry of T

(Earman [1989], p. 46)

appear as analytic truths.

And indeed, motivationalists have so far not adequately contended with the pressure a symmetry-relation between two models puts on the viability of regarding them as inequivalent. Read and Moller-Nielsen argue that doing so is admissible since "[t]here is nothing, after all, obviously absurd about admitting in principle undetectable facts into one's ontology" (Read and Moller-Nielsen [2020], p. 92). But this understates the problem with countenancing surplus structure as a candidate for inclusion into one's ontology. Surplus structure is not merely undetectable, it can even be *dynamically inert*. In that case, it is intangible, elusive, diaphanous; it doesn't latch onto the rest of reality in any way. This has been forcefully argued by Myrvold [2019], against the background of the dynamical view. He goes so far as to contend that for this reason, the words and mathematical objects by which we ostensibly denote surplus structure, such as 'standard of rest', perforce fail to receive a physical interpretation, so that discourse involving them is meaningless. From this he concludes that Earman's SP1 is not merely correct, but an analytic truth.

Two responses to Myrvold are available to the motivationalist. The first appeals to the inextricability of surplus structure. Myrvold's example of dynamically inert surplus structure is that of a standard of rest. With this example however, he stacks the deck in his favour since on the now standard formulation of Newtonian Gravitation Theory, the standard of rest can be neatly separated from the dynamically efficacious structure. Models of NGT have the form

$$\langle M, t_{ab}, h^{ab}, \nabla_a, \sigma^a, \varphi, \rho \rangle \tag{3.1}$$

where M is a four-dimensional manifold, t_{ab} and h^{ab} are the temporal and spatial metric, respectively, ∇_a a connection, σ^a a vector field, and φ and ρ scalar fields. The standard of rest is expressed by σ^a. Crucially however, there are ways of defining the dynamics of NGT according to which σ^a plays no dynamical role, so that one can simply delete it and is still left with a well-defined theory – the spacetime on which that theory is set is called Galilean spacetime. (Read and Moller-Nielsen [2020a]).

In this respect however, the standard of rest of NGT is not representative of other forms of surplus structure. In other cases, surplus structure does not consist in the fact that some bit of structure has been added on top of the dynamically efficacious structure, but in that said structure is subject to a choice of gauge, say. In such cases, the surplus structure is inextricable from the dynamical structure.[25] One example of this is the vector potential of electrodynamics. To make this point more precise, let me introduce the theory of electrodynamics in some of its formulations in more detail: Standardly, electrodynamics is formulated in terms of an invariant quantity called the Faraday tensor F_{ab}, a two-form on Minkowski spacetime meant to represent the electromagnetic field. The KPMs of Faraday tensor electrodynamics are of the form $\langle M, \eta_{ab}, F_{ab}\rangle$ where M is a manifold diffeomorphic to $\mathbb{R}^4$ and η_{ab} is a flat and complete metric, defining $\langle M, \eta_{ab}\rangle$ as Minkowski spacetime (Weatherall [2016a]). Minkowski spacetime comes equipped with a derivative operator ∇_a. The DPMs of the theory are singled out by Maxwell's equations

$$\nabla_{[a}F_{bc]} = \mathbf{0} \tag{3.2}$$

$$\nabla_a F^{ab} = J^b \tag{3.3}$$

Here, the first equation expresses that the Faraday tensor is closed and the second equation relates the Faraday tensor to the charge-current density J^a, represented by a vector field.

Alternatively, one can define electrodynamics as a theory of a vector potential (Weatherall [2016a]). In this version of the theory, the basic object is a one-form A_a representing the vector potential. The KPMs of the theory are of the form $\langle M, \eta_{ab}, A_a\rangle$, where $\langle M, \eta_{ab}\rangle$ is as before Minkowski spacetime. The DPMs are singled out by

$$\nabla_a \nabla^a A^b - \nabla^b \nabla_a A^a = J^b \tag{3.4}$$

The Faraday tensor is definable in this theory via $F_{ab} = \nabla_{[a}A_{ab]}$. It follows that the Faraday tensor associated with a vector potential must be closed. The vector potential is often considered a gauge quantity since the following transformation is a symmetry:

$$A_a \mapsto A'_a = A_a + \nabla_a \chi \tag{3.5}$$

Here, χ can be any smooth scalar field. What is crucial about this example is that unlike with the standard of rest, the surplus structure cannot be neatly separated out. We cannot present the theory in the form: Faraday tensor plus some independent piece of dynamically inert structure.

[25] See also Bradley [2021] for discussion of the potential inextricability of putatively surplus structure.

Dynamical and surplus structure are intertwined also in other theories – indeed, this is the default case. Examples such as the standard of rest of NGT are very much atypical. Indeed, there is also a formulation of NGT on which the standard of rest is not given in isolation but is "baked into" the spatial metric. Such formulations afford us a response to Myrvold's allegation that terms such as 'standard of rest' receive no physical interpretation. We can argue that absolute velocities and the standard of rest receive their physical significance from the dynamically efficacious metric from which they are definable. Myrvold's argument therefore fails to establish SP1 as analytic.

The second argument to be offered on behalf of the motivationalist appeals to the fact that structure that is at one point surplus may subsequently assume a dynamical role when the theory is modified, extended to new domains, or superseded and retained as a limiting case of a successor theory. The clearest example of this is the vector potential of electrodynamics. While often seen as surplus in classical electrodynamics, the vector potential arguably becomes indispensable once electrodynamics is integrated with quantum mechanics, as the Aharonov-Bohm effect shows. (see Belot [1998])

There are other examples to illustrate this phenomenon. For instance, Lorentz's ether theory requires a standard of rest, even though such a standard constitutes surplus structure in Newtonian Mechanics (see Bradley [2021]). Thus, Lorentz's theory revives a bit of structure we had at one point excised from our theories. Yet another example are relativistic extensions of Bohmian mechanics, which require a preferred hypersurface of simultaneity not found in Minkowski spacetime, and hence absent from electrodynamics and special relativity (see Dürr et al. [1997]). What this shows is that structure that in one theory appears surplus may yet be assigned a dynamical role in subsequent theorizing. Contra Myrvold then, there can be no presumption that all apparently surplus structure is indeed unphysical.

4.2 Moderate interpretationalism

Even if, contra Myrvold, we think there is nothing incoherent about believing in structure variant under symmetries, we might still think that the default should be to interpret symmetry-related models as equivalent. That is the view defended by Luc [2023]. As Luc

points out, many motivationalist arguments are directed against, and effective against, only a dogmatic, flat-footed form of interpretationalism, according to which declaring symmetry-related models equivalent is invariably justified, end of the story. Luc notes that there are more subtle versions of interpretationalism, according to which one has fairly strong initial justification for interpreting symmetry-related models as equivalent, which however can be defeated by subsequent developments such as a failure to find an adequate reformulation of the theory. Moreover, since there are positive arguments for interpreting symmetry-related models as equivalent, it is not obvious that an initial interpretationalist attitude is less cautious than a motivationalist one, despite what motivationalists maintain. While motivationalists often speak as though the rational default attitude towards symmetry-related models is a judgment of inequivalence, they have done too little to justify their claim that of the three initial, preliminary attitudes one might adopt towards symmetry-related models – interpretation as inequivalent, interpretation as equivalent, and suspension of belief –, the first is best.

Luc's point is valid of course. One possible but somewhat cowardly reaction motivationalists might opt for is to concede the point and either retreat to the view that the default attitude towards symmetry-related models should be agnosticism, or that the rational attitude will depend on the case, but will invariably be either belief in inequivalence or agnosticism. Another, even weaker position for motivationalists to take is that a default judgment of equivalence may sometimes be the best option but will have to be uncertain and tentative – full confidence in the equivalence of symmetry-related models can only be obtained via reformulation. That view however can hardly be called motivationalist, and already coincides (at least in the relevant cases) with what Luc calls 'graded' and 'concessive' interpretationalism – views according to which one is to make an initial judgment of equivalence, which however can be undermined by failures to find an adequate reformulation of the theories in question.

If we wanted to defend a more full-bodied form of motivationalism, we would have to make the case for a blanket presumption of inequivalence in cases of symmetry-related models. One way to do so might be to appeal to a broad commitment to scientific realism: we ought to believe whatever our current best theories say, even if they posit symmetry-variant structure. But this appeal rings empty and dogmatic. As ever, one man's modus ponens is another man's modus tollens. And in any case, the interpretationalist's position is not that

theories with symmetries are completely detached from reality but merely that some of their structure must be gauge – a view that is not obviously threatened by the best arguments for scientific realism, and indeed a view that is a form of "selective" realism, the like of which has recently been garnering much support.[26] So, a flat-footed appeal to scientific realism will not suffice to vindicate the motivationalist's position. Ultimately, what I think motivationalists should say is that the correct initial attitude to take towards a theory's symmetry-invariant structure will depend on a variety of factors, but that in many cases an initial judgment of inequivalence is indeed the most reasonable position and that even in circumstances in which there is a lot of pressure not to view a theory's symmetry-variant structure as physical, the case is normally not clear-cut.

Let us first look at two cases in which an initial judgment that symmetry-related models are inequivalent seems eminently reasonable. The first are the Galilean symmetries of NGT. As we saw in Chapter 1, it was for the longest time completely mysterious what ontology could ground the equivalence of models related by a kinematic shift. Only the introduction of Galilean spacetime shed light on this problem, which had seemed intractable for centuries.

The second case is the scale symmetry of NGT.[27] This symmetry consists of a simultaneous scaling of the particle masses by a factor μ, the inter-particle distances by a factor λ, and the durations between events by a factor τ, subject to the constraint $\lambda^3 = \mu\tau^2$. Two factors make an initial judgment of inequivalence appear reasonable here. First, the symmetry in question has the air of being fairly artificial and contrived. It is constituted by an involved transformation featuring durations, lengths, and masses, and the side constraint looks convoluted. The symmetry is therefore a far cry from eminently natural symmetries such as Leibniz shifts or Galilean boosts. This might make one inclined to write the scale symmetry off as an oddity or a coincidence that does not carry physical significance in the way these more natural symmetries do. Secondly, and relatedly, it will initially be far from clear how to reformulate the theory in a way that renders models related by the scale symmetry isomorphic. It will be less clear still whether such a reformulation will yield a theory that can match NGT in simplicity and elegance. All in all then, one's initial reaction to the scale symmetry might and arguably should be that the models it relates are inequivalent. In saying this, I am not claiming that this initial verdict is ultimately correct – indeed, as I will

[26] For defences of selective realism, see inter alia Harker [2013], Benitez [2019]
[27] For detailed discussion of the scale symmetry, see Jacobs [2022b], Dewar [2024], and Chapter 4 of this thesis

argue in Chapter 4, we can reduce NGT with respect to this symmetry. Dewar [2024] and Jacobs [2022b] have also provided sophistications of NGT with respect to scale. Nonetheless, the scale symmetry remains a case in which an initial judgment of equivalence would have been much too hasty.[28] A relevantly similar example which warrants attention but which unfortunately cannot be discussed at length in this thesis is the so-called CPT-symmetry.[29]

All that being said, motivationalists should concede that in certain cases, there is substantial pressure not to regard symmetry-variant structure as physical. In consequence, it may in these cases well be more reasonable than not to interpret symmetry-related models as equivalent. A good example of this is electrodynamics in the vector potential formulation. As a result of its gauge-symmetry, the vector potential has a really very unpleasant property: its time evolution is not uniquely fixed by the dynamics of the theory (Belot [1998]). This makes a reification of the vector potential an unattractive option even at the outset. What is noteworthy however is that even this case is not entirely clear-cut. Alternative interpretations of the theory, for instance in terms of the Faraday tensor, face their own problems. Principal among these is that such an interpretation cannot explain why the Faraday tensor should be closed (see Dewar [2019], Jacobs [2024]). Moreover, there are ways around the problem of indeterminism, principal among which is gauge-fixing (Maudlin [2018]). So, even in these cases the motivationalist is right to insist that clarity ought to be achieved by reformulation.

To situate my view relative to the spectrum of attitudes sketched by Luc [2023] then – ranging from steadfast interpretationalism to full motivationalism via graded and concessive interpretationalism, I want to say the following: how much pressure a symmetry puts on us to interpret the models it relates as equivalent is not uniform, but varies from case to case. Some symmetries invite judgments of equivalence more strongly than others. However, in every case known to me, the case for a judgment of equivalence *ab initio* is too weak to be compelling. Within Luc's classification then, I remain a motivationalist, albeit one that is open to the possibility that in certain specific cases, an attitude of graded interpretationalism might be the correct one to take.

[28] This case is also discussed in Read [2025]
[29] For discussion, see e.g. Earman [1989], Pooley [2002], and Saunders [2007]

4.3 External sophistication

Even if motivationalists are right to insist on reformulation, interpretationalists may argue that the demand for a replacement theory is trivial to satisfy. Wallace for instance insists that we possess "powerful, general resources by which structure can be subtracted from mathematical theories" (Wallace [2022b], p. 336) But while correct mathematically, this fails to refute the motivationalist position. Even if it is virtually guaranteed, as a matter of general mathematical fact, that a theory can always be reduced in a way that eliminates all symmetry-variant structure, there is no assurance that the resulting theory will match the original in elegance and explanatory strength.

Another idea for reformulating theories in a "cheap" and readily available way is to replace a theory exhibiting troublesome symmetries with what in the literature is called its "external sophistication" (see Dewar [2019], Martens and Read [2021]). Roughly, a theory's external sophistication is supposed to be the theory that corresponds to treating symmetry-related models of the original theory as isomorphic. To make this notion more precise, we need a modicum of category theory.[30]

First, a *category* C is a collection ob(C) of objects and, for each pair of objects $a, b \in ob(C)$ a set hom (a, b) of morphisms, or arrows, from a to b. For all $a \in ob(C)$, there exists an identity morphism $1_a \in$ hom (a, a). Any category C comes equipped with a binary operation $\circ$, the composition of morphisms. $\circ$ is associative and for any morphism $f \in \text{hom}(a, b)$, we have that $f = 1_b \circ f = f \circ 1_a$.

Categories bring with them their own standard of equivalence. Thus, an equivalence of categories is given by what is called an equivalence functor. A functor is a mapping from one category to another. More precisely, if C and D are categories, a functor $F: C \to D$ is given by an object $F(a) \in ob(D)$ for every object $a \in ob(C)$ and a morphism $F(f): f(a) \to f(b)$ for every morphism $f \in$ hom (a, b), subject to the constraints that for all $a \in ob(C)$, $F(1_a) = 1_{F(a)}$ and that for all morphisms f, g, $F(f \circ g) = F(f) \circ F(g)$.

[30] The category theoretic way of understanding external sophistication is by now standard and has been advanced e.g. by March [forthcoming]

An equivalence functor between two categories C and D is now a functor that has the properties of being full, faithful, and essentially surjective:

- Fullness: for all morphisms $g \in \hom(F(a), F(b))$, there exists a morphism $f \in \hom(a, b)$ such that $g = F(f)$
- Faithfulness: for any two morphisms f, g such that $f \neq g$, $F(f) \neq F(g)$
- Essential surjectivity: each $d \in ob(D)$ is isomorphic in D to an object d' such that $d' = F(a)$ for some $a \in ob(C)$

The strategy of external sophistication is articulated against the background of a category theoretic way of looking at theories.[31] According to this approach, theories are to be understood as associated with categories, and the appropriate standard for theoretical equivalence is categorical equivalence. While one can (and might want to, as we will see) associate a number of distinct categories with a theory T, there is a canonical choice $C(T)$: in *C(T)*, the models of T are the objects, and fully structure preserving maps between models are the arrows. In what follows I will take it as read that when I speak of the category associated with a theory, I mean the canonical category unless otherwise specified.

Equipped with the idea of the category $C(T)$ of a theory T, we can now get a handle on the idea of external sophistication. Roughly, the idea when dealing with a theory with symmetries is to start with the canonical category associated with the theory and then to add arrows in such a way that any two symmetry-related models are isomorphic in the resulting category (see e.g. Dewar and Eva [unpublished][32]). Slightly more precisely, we can identify a theory T's external sophistication T_S with whatever theory possesses a category $C(T_S)$ that differs from the original theory's category C(T) precisely in that all symmetry-related models are related by an isomorphism (in $C(T_S)$).[33] In other words, one can think of external sophistication as a procedure that involves first enriching the original theory's category $C(T)$ with isomorphisms for every pair of symmetry-related models and then taking as one's new theory *whatever theory it is* that possesses this new category $C(T_S)$.

[31] This approach was first introduced by Weatherall. See e.g. Weatherall [2016] and Rosenstock et al. [2016]. For a critical discussion of the categorical approach see also Weatherall [2021]

[32] This particular attempt at fleshing out external sophistication suffered from mathematical flaws and the paper was subsequently withdrawn from circulation (Dewar, personal communication). Below, I identify a range of technical issues that any account of external sophistication must overcome. I am sceptical about the prospects of doing so.

[33] Importantly, we here mean isomorphisms in the category. The models related by these isomorphisms need however not be isomorphic in the set-theoretic sense!

For a category $C(T_S)$ to count as a sophistication of a category $C(T)$ it must meet the following condition:

The S-I (symmetries as isomorphisms) Condition: $C(T_S)$ must satisfy

- a is an object of $C(T_S)$ if and only if a is an object of $C(T)$
- If a and b are objects of $C(T)$ and $f \in \hom_{C(T)}(a, b)$, then $f \in \hom_{C(T_S)}(a, b)$
- If a and b are objects of $C(T)$ representing models M and N of T respectively, and if M and N are related by a symmetry of T, then there exists $i \in \hom_{C(T_S)}(a, b)$ such that i is an isomorphism

One way to capture this requirement is due to Geroch (see Weatherall [2021]) It makes use of the category theoretic notion of natural isomorphism. Let us first define natural transformations: given functors $F, G: C \to D$ a natural transformation η from F to G is a family of morphisms $(\eta_x)_{x \in C}: F(x) \to G(x)$ such that for all $f \in C$, $\eta_{f(x)} \circ F(f) = G(f) \circ \eta_x$. If all of the η_x are isomorphisms then η is called a natural isomorphism. With this definition, Geroch's proposed criterion becomes:

(S1) Any auto-equivalence F of $C(T_S)$ that preserves empirical content is naturally isomorphic to the identity functor I: $C(T_S) \to C(T_S)$

But does external sophistication give us a trivial way to reformulate theories with symmetries? This idea has been criticized at length by Martens and Read [2021]. They focus mainly on the standard motivationalist point that declaring models to be equivalent when they are not isomorphic leads to opaque and potentially even indeterminate metaphysical commitments. While I am largely in agreement with their criticisms, I want to add some of my own. Martens and Read focus on the metaphysics behind external sophistication. I on the other hand want to highlight some technical problems with this programme. My criticisms are to be seen as a complement to, not a substitute for, Martens and Read's.

As a way to meet the motivationalist's demand for a replacement theory, external sophistication is so far only a promissory note. Moreover, as I will argue now, I doubt that the proposal can be fleshed out in a satisfactory way, since it faces a number of problems that seem difficult to overcome. External sophistication is a very coarse tool over which we

have little control. This manifests in problems for the proposal that may well prove insurmountable.

First, there is a technical problem with external sophistication: as it stands, the proposal does not determine a unique category $C(T_S)$ to be associated with a putative successor theory. Slightly more precisely, the proposal gives us a sufficient condition for external sophistication, viz. that any two symmetry-related models should be related by an isomorphism in the category $C(T_S)$. But in general, there will of course be many ways to enrich a category in a way that satisfies this requirement. Proponents of external sophistication must tell us precisely how to specify the category that we are to regard as the external sophistication of a theory.

A natural suggestion on behalf of external sophistication is that one should add, very roughly, as many arrows as is necessary at a minimum to satisfy the S-I condition but no more. In fact, one will be driven very quickly in this direction once one appreciates another problem for external sophistication. This problem is that in enriching a theory's category in the way required for external sophistication, a guarantee is needed that one has not inadvertently subtracted too much structure. After all, one wants the replacement theory to recover the dynamics of the theory one started with. This can only be achieved however if the new theory posits enough spacetime structure to support the dynamics. In consequence, one cannot add so many morphisms to the theory's category as to erase so much structure that the new theory violates Earman's symmetry principle (SP2) (see above).

The proposal at hand is therefore this: Choose for $C(T_S)$ the smallest enrichment of $C(T)$ that satisfies (S-I). But this suggestion is far from unproblematic. For one, as stated in this simple form, it involves an incorrect uniqueness assumption. In general, it is not the case that for any theory T and category $C(T)$ there should be a unique smallest enrichment by isomorphisms that could be treated as the theory's sophistication. To see this, consider the following example.

Let C be a category with two objects, a and b. Let $\hom(a,b) = \emptyset$. Choose hom (a,a) so that the automorphism group of a is Z4. Choose hom (b,b) so that the automorphism group of b is V4, the Klein four-group. Now, let us sophisticate this category by declaring a and b

isomorphic. To do this consistently we must add arrows in such a way that hom (a, a) is isomorphic to hom (b, b), and hom (b, b) is isomorphic to hom (a, b). That means that a and b need to be assigned the same automorphism group - call it G. Given $Aut(a) = Z4$ and $Aut(b) = V4$, G must have both Z4 and V4 as subgroups. There are exactly two groups of order 8 that satisfy this requirement, either $G = D4$ (the dihedral group of order 8) or $G = Z4 \times Z2$. By Lagrange's theorem, each choice is minimal. Hence, there is no unique minimal sophistication of C, not even up to categorical equivalence.

Moreover, even if we had a principled way of choosing among a category's minimal enrichments for the purposes of external sophistication, that choice may well not in general correspond to the category we want to choose for $C(T_S)$, i.e., the category associated with the *internal* sophistication of T. By an internal sophistication is here meant a full mathematical reformulation of the theory that yields a theory in which any two SRMs are isomorphic (see the next chapter for more detailed discussion). This worry is not at all unsubstantiated. In many cases, if one internally sophisticates a theory so as to render its symmetry-related models isomorphic, there will be *many* isomorphisms relating the SRMs. Consider the example of Dewar's [2019] theory of handedness T_H, viz.

$$T_H = \{\forall x(Lx \vee Rx), \forall x \neg(Lx \wedge Rx)\} \tag{3.6}$$

Intuitively, the theory says that everything is either lefthanded or righthanded, and that nothing is both. The theory is not sophisticated, since the mapping L to R and R to L is a symmetry that relates non-isomorphic models. Here, the models are ordinary L-structures, where L is the first-order language with signature $\Sigma = \{L, R\}$. Dewar therefore introduces some machinery to sophisticate this theory (internally, by modifying the models). The new theory replaces the L-structures of the original theory with what Dewar calls "dehanded pictures". A dehanded picture M is made up of

- A domain D_M
- A two-element set A_M
- An interpretation function $I_M: A_M \to \mathcal{P}(D_M)$

Truth in a dehanded picture is defined in the first instance relative to a variable assignment and an assignment $V: \Sigma \to A_M$. For $\Pi \in \Sigma$ we have

$M \vDash_{V,v} \Pi x$ if and only if $v(x) \in I_M(V(\Pi))$

The clauses for complex formulae are the obvious ones. It is worth noting, albeit irrelevant for present purposes, that this semantics yields a determinate truth value for a formula φ just in case φ is logically equivalent to φ^*, where φ^* is the result of substituting in φ L for R and R for L (Dewar [2019]).

What is relevant to the discussion is the number of isomorphisms between symmetry-related models of the theory. So, we need to say what isomorphisms are in this case. First, a homomorphism h between dehanded pictures M and N is given by functions

- $h_1: D_M \rightarrow D_N$
- $h_2: A_M \rightarrow A_N$

such that for $a \in A_M, h_1(I_M(a)) = I_N(h_2(a))$. As is standard, an isomorphism is an invertible homomorphism. In the new theory, any two models related by the L/R symmetry are indeed isomorphic. However, there is no canonical isomorphism, and any two SRMs will in general be related by a multitude of isomorphisms. To see this, consider a dehanded picture M such that

- $D_M = \{a, b, c, d, e\}$
- $A_M = \{0, 1\}$
- $I_M(0) = \{a, b, c\}$ and $I_M(1) = \{d, e\}$

This dehanded picture is related to its mirror-symmetric counterpart obtained by permuting 0 and 1 by a total of 3!2! = 12 isomorphisms.

This leads to a second concern about the proposal to obtain $C(T_S)$ by adding as few arrows as possible to $C(T)$: if one does so, one has no guarantee that the morphisms one has added have a natural interpretation in terms of mappings between the base manifolds of the models one treats as objects in the category. This in turn suggests that there might not be a natural theory corresponding to the category $C(T_S)$ at all. Consider once again the example just given. The sophistication we know to exist is Dewar's internal sophistication. But in that theory, isomorphic models are in general related by many isomorphisms. It is far from clear what a theory with precisely one isomorphism for each pair of symmetry-related models would look like or if it exists at all. It certainly would not be as natural or attractive as Dewar's sophisticated theory.

These are serious challenges for the external approach to sophistication. But even if they can be overcome, there arises yet another difficulty: even if we can determine a unique category $C(T_S)$ to be associated with the sophistication T_S of a theory T, and even if $C(T_S)$ is the category we want in the sense that its arrows are amenable to interpretation, $C(T_S)$ still does not in general pick out a unique theory T_S as the sophistication of T. The reason for this is that in relevant cases, theoretically inequivalent theories will be associated with the same category, i.e., will be categorically equivalent. Dewar [2019] presents a range of cases in which some theory exhibiting symmetries can be reformulated in more than one way so as to eliminate the symmetry. The strategies available for this purpose Dewar labels reduction and sophistication. We will look at these strategies in detail in the next chapter. For now, it suffices to note that in some of the cases Dewar discusses, they lead to categorically equivalent but explanatorily inequivalent theories. The resulting theories are therefore theoretically inequivalent even though they induce the same category (up to categorical equivalence).[34] This is a much more devastating problem for the program of external sophistication than the technical problems discussed above since it points to an inherent expressive deficiency in the categorical representations of our theories.

That being said, there has been some recent work that seeks to remedy precisely this problem (see March [forthcoming]). The crucial idea is to represent in the category associated with a theory not only the DPMs of the theory but also the KPMs. This makes categorical equivalence a much more fine-grained standard of equivalence. As March shows, many theories that are categorically indistinguishable at the level of the DPMs come apart at the level of the KPMs. But it remains to be seen whether this suffices to vindicate categorical equivalence as a standard appropriate to physics, especially in light of the criticisms of Weatherall [2021]. And even if that is the case, still further work is needed to defend the programme of external sophistication against the criticisms raised here and in Martens and Read [2021].

This concludes my defence of motivationalism against outside foes such as interpretationalists. In the next section, I turn inwards. I will develop my own version of motivationalism, in opposition to other views held within the motivationalist camp.

[34] In the background here are the following commitments, which I will not defend but which strike me as compelling: Theoretical equivalence requires explanatory equivalence, and so mere categorical equivalence does not suffice for theoretical equivalence. (For what it's worth, I also have some doubts over whether categorical equivalence is necessary for theoretical equivalence.)

5 What is the correct version of motivationalism?

If what I have said so far is correct, motivationalism is broadly the correct attitude to take. But that still leaves many questions of detail open. It is these issues that we turn to next. In 5.1, I argue that we must reformulate theories in a way that eliminates their symmetries before we are permitted to interpret their symmetry-related models as equivalent. Here, I disagree with certain motivationalists who maintain that it suffices to delineate the ontology that grounds this equivalence without necessarily constructing a full reformulation. In 5.2, I investigate what conditions a reformulation must meet if it is to warrant a judgment that the original theory's symmetry-related models are equivalent. Jacobs [2022a] has argued that we ought to insist on "intrinsic" reformulations, by which he roughly speaking means theories all of whose mathematical structure purports to directly represent structure that is physical according to the theory. I argue against this requirement of intrinsicality.

5.1 In Defence of Reformulation

Is reformulation really necessary before we are permitted to regard symmetry-related models as equivalent? Even if the motivationalist is right to demand a specification of the ontology shared by such models, why can we not describe the ontology in a way that does not amount to full reformulation? Of course, we may want to say slightly more than the interpretationalist, who is content with committing himself to "whatever the invariants of the symmetries are". But, as Read [2022] points out, it does not follow that we cannot give a sufficiently clear picture of the shared ontology without reformulation.

Against this, Jacobs [2022a] argues that reformulation *is* necessary for an inference of equivalence between symmetry-related models. I agree with Jacobs that reformulation is necessary, but my reasons for thinking so are different from his. Jacobs' arguments are mainly directed against extrinsic reformulation, by which he roughly speaking means defining the replacement theory with the aid of structure not directly representational of physical structure. Examples of extrinsic definition are coordinate-based definitions, and constructions involving quotienting. Jacobs argues indirectly against the sufficiency of giving the ontology shared by symmetry-related models without reformulation, on the basis

that extrinsic reformulation gives you said ontology, but crucially not in explicit or perspicuous form. Thus, one way to reconstruct the outline of his argument is as follows

(1) Extrinsic reformulation of a theory T yields a theory T' with a well-defined ontology, comprising precisely the symmetry-invariant quantities definable in T
(2) However, extrinsic reformulation does not license the interpretation of symmetry-related models as equivalent
(3) *A fortiori*, delineating the ontology shared by symmetry-related models does not license the interpretation of these models as equivalent

Whatever one makes of this argument – I disagree with Premiss (2), as I will explain in detail in the next section – it is important to realize that it does nothing to address the position sketched by Read [2022], according to which one has to *identify* the invariants shared by symmetry-related models, rather than just constructing a theory committed to nothing beyond these invariants, in order to be licensed to regard these models as equivalent. So, to counter Read's view, we need a new argument. I think that such an argument can be given.

Why finding invariants is arguably not sufficient to license an inference of equivalence is that it falls short of showing that an adequate dynamics can be expressed in terms of these invariants. This is a non-trivial task, and there is no guarantee that it can be accomplished. One issue that can arise is that it is not clear how the symmetry-invariant quantities should feature in dynamical equations. Consider for instance the scale symmetry of Newtonian Gravitation Theory.[35] This symmetry consists in a simultaneous scaling of the inter-particle distances by a factor λ, the particle masses by a factor μ, and the temporal distances between events by a factor τ, subject to the constraint $\lambda^3 = \mu\tau^2$. One invariant might be the ratio of the particle masses times the durations squared to the distances cubed. But it is far from obvious how to incorporate this into the dynamics. So, the task becomes that of identifying a subset of natural invariants that suffice for formulating an adequate dynamics. Since the demands for naturalness and sufficiency pull in opposite directions, there is in general no guarantee that this problem is not overconstrained.[36] For example, fairly natural invariants of the scale symmetry include distance ratios and mass ratios – but these do not suffice to replicate Newton's laws.

[35] For detailed discussion of this symmetry, see Jacobs [2022], Dewar [2024], and Ch. 4 of this thesis.
[36] Dewar [2019] makes roughly this point.

A further reason to be skeptical that identifying symmetry-invariants suffices to provide the ontology common to symmetry-related models is that in many cases, the best account of their shared ontology comprises either symmetry-variant quantities, or quantities that are not naturally viewed as part of the original theory's ontology at all. Take for instance Newton-Cartan Theory (see Malament [2012]). Its ontology includes a dynamical, and therefore potentially curved affine connection – an entity not naturally viewed as part of the ontology of Newtonian Gravitation Theory. Moreover, it includes symmetry-variant quantities such as locations in spacetime.

Indeed, it is often thought that in many instances, an ontology consisting entirely of symmetry-invariant quantities is unsatisfying since it requires "cosmic conspiracies". (Maudlin [2007], Arntzenius [2012], Dewar [2019], Jacobs [2024]) A cosmic conspiracy arises when certain fundamental quantities exhibit patterns that make them look as though they were not fundamental after all, but derivative of other quantities. For a simple example, take features of the Faraday tensor in electrodynamics, specifically, the fact that the Faraday tensor is closed. If one treats the Faraday tensor as fundamental, that it should be closed is a brute constraint one must impose. The resulting theory however can offer no internal explanation of this constraint – such an explanation would have to appeal to the fact that the Faraday tensor is the exterior derivative of a vector potential. But that would require one not to treat the Faraday tensor as fundamental after all.

Thus, cosmic conspiracies arise because a theory treats quantities as basic when it would be best not to, since doing so mystifies patterns among the putatively basic quantities that could be given a natural explanation. At the level of theory construction, this means having to impose rather *ad hoc* looking constraints in order to capture these patterns.

There are other well-known examples of conspiracies. (see e.g. Jacobs [2024]) One that also arises in the context of electrodynamics is that of holonomies. Roughly, these are integrals of the vector potential along closed loops. Treating these as fundamental requires that one impose the following constraint:

$$H(i \circ j) = H(i)H(j) \tag{3.7}$$

where $\circ$ denotes the concatenation of loops. Once again, this connection can be explained by appeal to the vector potential.

A further example, also discussed by Jacobs, is that of absolute masses and mass ratios. Mass ratios between particles have to be made to obey this principle:

$$r(a,b) = r(a,c)r(c,b) \tag{3.8}$$

which of course trivially holds when particles are assigned absolute masses. Finally, there is the case of the triangle inequality: in theories that treat inter-particle distances as fundamental, one must demand that they satisfy the following:

$$d(x,z) \leq d(x,y) + d(y,z) \tag{3.9}$$

This inequality is readily explained if one treats distances not as fundamental, but as relations between positions in space (Jacobs [2024]).

Crucially for present purposes, it is ontologies that treat symmetry-invariant quantities as basic that frequently give rise to cosmic conspiracies, as these examples illustrate – the Faraday tensor is an invariant of the gauge symmetry of electrodynamics, as are the holonomies; mass ratios are an invariant of mass scalings (see Martens [2022]), inter-particle distances are an invariant of Leibniz shifts. Indeed, Jacobs [2024] identifies some fairly general conditions under which treating symmetry-invariant quantities as basic gives rise to cosmic conspiracies. So, *pace* the proposal sketched by Read [2022], merely identifying the quantities invariant under a symmetry does not license us to regard symmetry-related models as equivalent. One must reformulate the theory, either by giving a dynamics in terms of the symmetry-invariants, or in terms of other objects.

* * *

Having defended the need for reformulation, let us now turn to the question of what conditions a reformulation must meet if it is to warrant the judgment that the original theory's symmetry-related models are equivalent. We will consider two conditions that have been defended in the motivationalist literature: first, Jacobs' requirement of intrinsicality (Section 5.2), and secondly, the requirement that the reformulated successor theory must give us a "metaphysically perspicuous characterisation" of the ontology that underpins the equivalence of the symmetry-related models (Section 5.3) – call this the perspicuity requirement for short.

What are intrinsicality and perspicuity? Jacobs defines a theory to be intrinsic if in formulating it we "lay down a set of relations, functions and operators which explicitly

represent the world's physical structure." (Jacobs [2022a], p. 13). Perspicuity on the other hand does not have a universally accepted definition in the literature and is no doubt understood differently by different authors. Let me therefore say more about what I take metaphysical perspicuity to be. I take a theory to be perspicuous just in case we have a clear grasp of what, according to the theory, the world is like. This will include a grasp of the theory's ontology, i.e. of what, according to the theory, the constituents of reality are. But it will also include for instance an understanding of whether, according to the theory, the world is governed by certain principles of locality or separability, say. In general, the more questions we are in a position to answer about the metaphysical picture of the world a theory offers, the more perspicuous that theory is to us. I therefore take perspicuity to be a matter of degree. The paradigm of a theory that at the time of its conception was maximally imperspicuous is of course quantum mechanics. But even theories that are comparatively well understood can raise residual interpretive questions. For an example, take the status of Killing fields in General Relativity (see Curiel [2018]). As Curiel argues, it is a difficult open question whether they should be taken to represent something physical. Perspicuity is furthermore linked to explanation. The more perspicuous a theory is, the better we understand its explanatory capacities. Perspicuity allows us to assess the quality and scope of the explanations a theory has to offer.

Perspicuity is something that can be achieved through interpretation. In Latin, *perspicere* means 'to see through'. The goal of interpretation is to see through a theory and to understand what it says about the world. How perspicuous a theory is therefore depends on historical, sociological, and psychological factors. A theory can be imperspicuous at one point in time and perspicuous later on. Take the example of the theory of Einstein Algebras (EA). Initially, this theory was not well-understood. It was thought that EA differed in important respects from the theory of General Relativity. Through interpretive work however, we have come to understand that EA is equivalent to General Relativity (Rosenstock, Barrett and Weatherall [2015]). We have made EA perspicuous (or at least as perspicuous as General Relativity). A hypothetical similar example is sketched by Wallace and Timpson [2012]. They ask us to imagine we were given a formulation of classical mechanics on phase space without the knowledge of how this theory relates to ordinary n-particle classical mechanics on Euclidean space. One might initially think that the theories differ since the phase space theory appears to describe the evolution of a single particle in an extremely high-dimensional space. Interpretive work, part of which would involve

uncovering the mathematical correspondence between the two theories, however would then reveal them to be fully equivalent. This would be a way of making the phase space theory perspicuous.

Undoubtedly, intrinsicality and amenability to perspicuous interpretation are desiderata for reformulation, and for theory construction more generally. Nonetheless, I will argue that judgments of equivalence for symmetry-related models can sometimes be warranted even by reformulations that are neither intrinsic nor perspicuous. The criteria of intrinsicality and perspicuity are therefore in general overly demanding. However, in saying that these requirements need not invariably be satisfied I am not arguing that they should be abandoned without replacement, or that just any reformulation of a theory with symmetries will suffice to convince us that the original theory's symmetry-variant structure is surplus. My aim is rather to isolate precisely what conditions a reformulation must satisfy if it is to warrant a judgment of equivalence for the symmetry-related models of its predecessor. What truly matters, I will argue, is that the reformulated theory can be recognised to be not explanatorily deficient relative to the original theory. This condition – call it the *recognisable adequacy condition* – is in essence just that there must be a way to ascertain the absence of defeaters of the reformulation, such as that the reformulated theory lacks explanatory strength, etc.

In my view, a theory's being perspicuous is sufficient for us to be able to recognise whether it is adequate. Sometimes however, we can see that a theory is an adequate reformulation even before we have a perspicuous characterisation of the metaphysical picture of the world it offers. These cases are arguably quite rare. I take it that the default way of judging whether a reformulation is adequate is to first interpret it so as to render it metaphysically perspicuous and to then adjudicate its adequacy on the basis of that interpretation, in line with the procedure envisioned by defenders of standard motivationalism. Nonetheless, there are important exceptions. This is what I will argue in Section 5.3. First, I will argue against a requirement of intrinsicality.

5.2 Against intrinsicality

Jacobs [2022a] has defended the view that in order to be licenced to interpret symmetry-related models as equivalent, we need not only a reformulation of the theory, but an *intrinsic* reformulation. To formulate a theory intrinsically is to "lay down a set of relations, functions and operators which explicitly represent the world's physical structure." (Jacobs [2022a], p. 13) What Jacobs defends then is the following requirement:

> *Requirement of Intrinsicality*: A reformulation T' of a theory T only licenses an interpretation of symmetry-related models of T as equivalent if T' is formulated intrinsically.

To illustrate the notion of intrinsicality, Jacobs [2022a] provides us with two examples of theories that fail to be intrinsic:

1. A version of NGT set in Galilean spacetime in which the base manifold M is equipped with Galilean structure via quotienting of coordinatisations of M by the equivalence relation x ~ x+vt
2. Earman's definition of Maxwell Gravitation in terms of an equivalence class of affine connections

Let me give some more detail on the second example: Defined extrinsically, the models of Maxwell Gravitation are of the form $\langle M, t_a, h^{ab}, [\nabla], T^{ab} \rangle$, with M a four-dimensional base manifold, t_a and h^{ab} compatible temporal and spatial metrics, respectively, T^{ab} the Newtonian mass-momentum tensor, and $[\nabla]$ an equivalence class of rotationally equivalent flat derivative operators, rotational equivalence of ∇ and ∇' here captured by the condition that $\nabla^{[a}\eta^{b]} = 0$ just in case $\nabla'^{[a}\eta^{b]} = 0$ for all unit timelike vector fields η^a (Dewar [2018]). The equations singling out the DPMs of Maxwell Gravitation make reference to the mass density field $\rho = t_a t_b T^{ab}$ and demand that at any point at which ρ does not vanish, the following hold:

$$t_a \nabla_n T^{na} = 0 \tag{3.10}$$

$$\nabla_m(\rho^{-1}\nabla_n T^{nm}) = -4\pi\rho \tag{3.11}$$

$$\nabla^c(\rho^{-1}\nabla_n T^{na}) - \nabla^a(\rho^{-1}\nabla_n T^{nc}) = 0 \tag{3.12}$$

for any $\nabla \in [\nabla]$.

In each case, it is the fact that the objects involved i.e., the coordinate functions and connections, respectively, do not themselves represent physical structure but have an auxiliary function that marks the theory as extrinsically defined. One can contrast each of these extrinsic formulations with intrinsic formulations of the same theory. We already saw an intrinsic formulation of NGT on Galilean spacetime above. For Maxwell Gravitation, Chen [2023] and March [forthcoming], building on work by Weatherall [2018a], have provided an intrinsic definition, directly in terms of a standard of rotation $\circlearrowleft$.

Now, it would be difficult to deny that intrinsic definitions of theories are especially valuable. They are an ideal we should aim for. But can we only be justified in interpreting certain symmetry-related models as equivalent if we possess an intrinsic reformulation of the theory in which they occur? I will argue that the answer is no. As I hope to show, Jacobs' requirement of intrinsicality is unreasonably demanding.

I will proceed as follows. I will first survey various arguments for intrinsicality. I will argue that they fail but that one of them – an argument due to Jacobs [2022a] – contains the seeds for a novel, more compelling argument for intrinsicality. I will call this argument the 'hidden flaws' argument. In brief, the idea behind the argument is that non-intrinsic theories can hide flaws, so that we cannot say with certainty that they have the same explanatory power as the theories they are constructed to replace. Since we cannot rule out that they are explanatorily deficient then, non-intrinsic reformulations do not provide a warrant for interpreting symmetry-related models of their predecessors as equivalent.

Having isolated what I take to be the strongest currently known argument for a requirement of intrinsicality, viz. the 'hidden flaws' argument, I finally argue that it, too, fails – at least insofar as it is meant to establish a requirement of intrinsicality. Let us therefore look at possible arguments for the requirement of intrinsicality. One argument has it that extrinsically reformulated theories do not show the structure they seek to eliminate to be dispensable.[37] After all, if a theory is formulated extrinsically, there may be no way to express the dynamical laws of the theory without appealing to geometric objects that are at least partly redundant.[38] To illustrate this point, consider how one would go about formulating a theory of gravitation on Maxwell spacetime, characterized extrinsically in

[37] This argument was suggested to me by Adam Caulton
[38] March [forthcoming] also discusses such cases

terms of [∇]. What one has to do is lay down equations that define the DPMs of the theory one is after, Maxwell Gravitation. One such equation is the following (Dewar [2018], p. 257):

$$\nabla_a(\rho^{-1}\nabla_n T^{na}) = -4\pi\rho \tag{3.13}$$

But note that this equation makes reference to a connection ∇ from [∇], although it is independent of the choice of representative from the equivalence class [∇]. This however raises the worry that the connection is indispensable to the theory, so that Maxwell Gravitation, at least in its extrinsic formulation, cannot provide grounds for rejecting belief in a standard of linear acceleration. The correct response here is to think of the dynamical equations merely as a constraint which permits a compact definition of the space of the theory's DPMs. While reference to a connection may be practically indispensable then, the connection merely plays an auxiliary role. The theory's commitments ought to be read off the internal definable structure of its models, not off the DPMs.

One might hope to construct a different argument for the intrinsicality requirement on the basis of March's work on the topic. March [forthcoming] in the first instance offers an account of the value of intrinsic formulation. She points out that if a theory is formulated extrinsically, it will in general be non-trivial to determine whether an equation in the language of the theory is meaningful or not. Take for instance once more the example of Earman's definition of Maxwell spacetime. As we just saw, he defines it in terms of the equivalence class [∇] of flat connections that agree on a standard of rotation. The language of the theory therefore contains (expressions for) all of the elements of the equivalence class. However, since the theory is only committed to the structure common to the elements of the equivalence class, an equation in the language of the theory that involves a connection ∇ will only be meaningful if any other connection $\nabla' \in [\nabla]$ is substitutable for ∇ *salva veritate,* a condition which will have to be verified before an equation can be used in calculation or argument.[39] In theories that are formulated intrinsically on the other hand, any syntactically well-formed equation will automatically be meaningful. March points out that this captures a sense in which intrinsically formulated theories are more perspicuous than their extrinsic equivalents.

[39] It is worth noting that this requirement is necessary but not sufficient for being meaningful – this is discussed at greater length in March (unpublished). It has no relevance to the present discussion and so I will bracket it.

March does not discuss whether the relatively lesser perspicuity of extrinsically formulated theories is sufficient to establish the requirement of intrinsicality, so this task will fall to us. My impression is that it is not. It would be very difficult to argue that only theories expressed in a language in which being meaningful coincides with being syntactically well-formed are perspicuous enough to guide our reasoning and beliefs. English for one is a language in which all sorts of nonsense can be expressed. Think about the liar sentence for instance, or take Chomsky's sentence 'Colourless green ideas sleep furiously' (Chomsky [1957], p. 15). Nonetheless, virtually all of our theorizing proceeds in English and, occasional linguistic confusions notwithstanding, may be regarded as generally reliable. There are no general grounds then on which one ought to suspect that the imperspicuity a theory may incur due to the possibility of well-formed but meaningless expressions could support the requirement for intrinsicality.

Of course, it might be that there is something specifically about the methods of extrinsic formulation that makes it prone to engendering error and confusion. It must be granted that when working in an extrinsically formulated theory, care must be taken to ensure that the equations one uses are meaningful, which means verifying that they are independent of the choice of representatives from equivalence classes. But while this may make it less readily apparent which expressions of the language are sound and admissible, in cases of doubt certainty can always be established through straightforward, albeit perhaps cumbersome calculation.

Jacobs [2022a] provides different arguments for the requirement of intrinsicality. He faults extrinsically formulated theories for allegedly (a) failing to provide causal explanations of the phenomena they describe, (b) assuming the representational equivalence of symmetry-related models as a brute fact, and (c) involving physically inert structure in their account of reality. Of these, (c) is the principal charge.

In what is to follow, I will bracket (a) and (b). March [unpublished] argues convincingly that these criticisms do not get off the ground since Jacobs conflates extrinsic reformulation with external sophistication. However, (a) and (b) are effective against external sophistication at best. That leaves (c) to be addressed. Jacobs' claim is that on account of involving physically inert structure, extrinsic reformulations fail to deliver the perspicuous characterization of the ontology common to symmetry-related models that motivationalists

claim we need in order to be licenced to interpret them as equivalent. Thus, Jacobs' argument zeroes in on a requirement of perspicuity which extrinsic reformulations are alleged to fail. He states that

> On the view I have defended, we are only warranted to interpret SRMs as physically equivalent when we have both a perspicuous account of the theory's ontological commitments, and a perspicuous formalism from which we can 'read off' these commitments. The latter amounts to an intrinsic theory in the sense of Field (1980).
> (Jacobs [2022a], p. 15)

So, the main thrust of his argument is that extrinsic reformulations of a theory do not underwrite an interpretation of symmetry-related models as equivalent because they fail to give a perspicuous characterization of the ontology common to such models, or at least of the dynamics of how the objects in the theory's ontology evolve and interact. But this raises the question: why exactly should that be disqualifying? Of course, motivationalists often claim that a reformulation must offer a metaphysically perspicuous characterisation of the ontology common to symmetry-related models if it is to warrant the judgment that they are equivalent. But why should anyone think this? If the motivationalist standpoint is to be more than dogma, an argument is needed. Only in conjunction with such an argument could Jacobs' claim that extrinsically formulated theories fail to be perspicuous represent a complete case in favour of the requirement of intrinsicality.

Let us therefore attempt to supply an argument of this kind. The argument I want to suggest builds on the insights of March and Jacobs, particularly the link between intrinsicality and perspicuity they emphasize, but develops a more precise sense in which the lack of perspicuity that an extrinsically formulated theory may exhibit can mean that the theory is not apt to warrant an anti-realist attitude towards the ontological posits of its predecessor. I call it the 'hidden flaws' argument. Of course, since I intend to argue against both the intrinsicality and the perspicuity requirement, I believe that the 'hidden flaws' argument fails to establish a general need for either. Nonetheless, it gets at something important. Ultimately, what I take it to support is the 'recognisable adequacy' condition. But more on this later. Let us first put the argument on the table before we investigate what exactly it shows.

The 'hidden flaws' argument is this. What marks a theory as extrinsically formulated is that it does not present the structure it posits as physical in a closed, self-contained manner. Rather, auxiliary structure is invoked. A consequence of this is that an extrinsic

formulation does not without further work give us a clear handle on the structure it treats as basic. In particular, we do not in general know what a free-standing, pure definition of that structure looks like. But crucially, without such a grasp, one cannot decide whether the structure in question *should* be treated as basic. The reason for this is that one is left without assurance that the theory is committed at the fundamental level to structure that admits a simple and unified characterisation. Such knowledge however is crucial if one is to be warranted in favouring the new theory over the old and so interpret symmetry-related theories of the old theory as equivalent.

To illustrate the argument and to see its force, consider the example of electrodynamics in the vector potential formulation. An extrinsically reduced version of the theory will be formulated in terms of equivalence classes [A] of vector potentials A related by a gauge transformation. At first glance, this may seem like a decent substitute for the original theory – certainly no obvious flaws can be detected by simply inspecting the theory as stated. But what is the structure this theory treats as basic? Plausibly, the theory is committed at the fundamental level to the holonomies of the vector potential, since it is the holonomies that exactly capture the gauge-invariant content of the original theory (see Healey [2007]).[40] In other words, the theory is plausibly a formulation of 'loop' electrodynamics. But as is well-known, there are excellent reasons not to treat holonomies as basic: to do so, one must posit as brute, unexplained fact that certain connections, viz. composition laws of the form

$$H(s \circ t) = H(s)H(t) \tag{3.14}$$

govern the theory's basic entities (Jacobs [2024]). Of course, in the vector potential formulation of electrodynamics, these composition laws have an explanation: they can be derived from the definition of the holonomies as integrals of the vector potential. This tells decisively against loop electrodynamics. That it should seem as though the relations between the holonomies obtain in virtue of the presence of a potential when in fact there is no such thing would constitute a massive "cosmic conspiracy" (Jacobs [2024], Arntzenius

[40] Some might think that the [A]-theory is instead the Faraday tensor formulation of electrodynamics "in disguise". After all, any two vector potentials related by a gauge transformation give rise to the same Faraday tensor. Even if this turned out to be correct, the example would still illustrate the 'hidden flaws' argument, since the Faraday tensor formulation also lacks explanatory strength: it cannot explain why the Faraday tensor is closed

[2012]). Loop electrodynamics fails to preserve the unificatory power of the vector potential theory.

The crucial point for present purposes is that this flaw in the reduced theory is concealed by extrinsic formulation. Loop electrodynamics formulated extrinsically looks deceptively simple. On the page, the only difference relative to the original theory is that we have replaced a vector potential with an equivalence class of vector potentials. The massive loss of explanatory power engendered by this transition is not at all discernible from the extrinsic formulation of the theory, and indeed the surface syntactic similarities between the two theories may trick one into thinking that it couldn't possibly have occurred. One may therefore think that the reformulated theory warrants abandoning the original vector potential theory when in fact it does not. Dasgupta points out, recall, that "we can draw the conclusion [...] [viz., that a given theory's relevantly 'variant' quantities are not real] only when we have the alternative theory in hand *and have shown that all else is equal*." (Dasgupta [2016], p. 853-4, cit. Read and Moller-Nielsen [2022b], p. 93, emphasis added) The argument against extrinsic formulation is then that it makes the task of showing that really all else is equal difficult if not impossible. Recall from Section 2 Martens' 'no defeaters' condition: a reformulation of a theory that eliminates symmetries can only support the conclusion that the original theory's symmetry-related models are equivalent if the reformulated theory has the same explanatory strength as the original. According to the 'hidden flaws' argument, which one could also call the 'hidden defeaters' argument, one can only be certain that the reformulated theory passes this test if it is defined intrinsically.

Does the 'hidden flaws' argument succeed? I believe it to be a powerful argument that teaches important lessons. But as I will argue now, it cannot establish a requirement of intrinsicality. Note a common thread among the arguments for intrinsicality: the 'hidden flaws' argument's heavy emphasis on perspicuity mirrors the central theme of March's and Jacobs' discussion. Hence, I want to suggest that many arguments for a requirement of intrinsicality are in the first instance arguments for a requirement of perspicuity and only mediately arguments in favour of intrinsicality. Implicit in the discussions of intrinsicality due to Jacobs and March, and also in the 'hidden flaws' argument, is a wariness of the interpretive pitfalls that come with imperspicuous theories. Intrinsicality appears in these discussions as the best, and perhaps only, way to avoid imperspicuity and is to be demanded primarily for this reason.

But as I will argue, the connection between intrinsicality and perspicuity is much less straightforward than is perhaps commonly assumed. This will give us two ways to push back against the 'hidden flaws' argument *qua* argument for an intrinsicality requirement. The first is to deny that extrinsic methods invariably yield imperspicuous theories. If this is right, and if perspicuity suffices for us to be able to assess the adequacy of a reformulation, this would tell against a general requirement of intrinsicality. Extrinsic formulation is permissible just in case it does not interfere with our ascertaining whether or not there are defeaters of the reformulated theory. The second response is more radical. It is to deny that a perspicuous account of the ontology shared by symmetry-related models is strictly necessary for judgments of equivalence. I think that both of these responses succeed. More precisely, it seems to me that in many cases, a theory's being formulated extrinsically does not get in the way of obtaining a perfectly clear sense of the theory's ontology, which in turn suffices for assessing its adequacy. And where it does get in the way, that in itself does not necessarily mean that the theory so formulated does not give us grounds for interpreting the symmetry-related models of its predecessor as equivalent, or so I will argue. In the present section, I will limit myself to establishing the less radical claim, that perspicuity does not require intrinsicality. That even imperspicuous reformulations can in certain cases warrant judgments of equivalence for symmetry-related models is something I will argue for in Section 5.3.

Of course, one will wonder how I can still call myself a motivationalist if I am going to deny that we (invariably) need a metaphysically perspicuous account of the ontology shared by symmetry-related models before we are free to interpret them as equivalent. The reason I am a motivationalist nonetheless is that I insist we must replace the theory exhibiting symmetries with a new theory in which these symmetries are eliminated and which is recognisably as explanatorily adequate as its predecessor before we can infer the equivalence of symmetry-related models. This is not to say that a perspicuous account of a theory's ontology is not worth having, of course. The claim is merely that judgments of equivalence may be legitimate even in the absence of such a picture, as long as the explanatory adequacy of the reformulated theory can be established.

But more on this in the next section. For now, let us develop a response to the 'hidden flaws' argument. The first point to be made is that a failure of intrinsicality does not invariably lead to imperspicuity. Ordinarily, it seems to me, an extrinsic definition of a

structure does not prevent us from obtaining a perfectly clear, perspicuous picture of that structure. This should not be all that surprising, given how ubiquitous extrinsic definitions are in physics. For a first example, note that conformal structure is commonly defined via an equivalence class of metrics (Pitts [2006]).[41] There is nothing imperspicuous about proceeding in this way. But then, since I take perspicuity to be sufficient for allowing us to assess whether a theory is explanatorily adequate, it follows that extrinsic theories can often be evaluated with regard to their adequacy. As can be seen from actual debates in the philosophy of physics, in many cases, the lack of intrinsic formulations is not an obstacle to interpretive debates around theories since the participants in the debates have a firm intuitive grasp of the relevant theories.

Take for instance the second of Jacobs' examples of extrinsic formulation, i.e., Maxwell spacetime. Its features are what matters in the debate over whether Maxwell Gravitation is theoretically equivalent to Newton-Cartan Theory (Saunders [2013], Wallace [2020], March [forthcoming]). As introduced by Earman [1989], Maxwell spacetime was defined extrinsically, in terms of an equivalence class of connections. For a long time, a full intrinsic formulation of Maxwell Gravitation was not known – Chen [2023] and March [forthcoming] was the first to construct one. Nonetheless, it is possible to get an intuitive handle on the structure in question. What helps with this is the usual informal commentary one can offer to describe the structure, such as for instance that Maxwell Spacetime is Leibnizian Spacetime with an added standard of rotation (Earman [1989], p. 31). Perhaps even more importantly, one can come to understand this structure by understanding its symmetries, symmetries being a tremendously powerful tool for visualization and reasoning. Indeed, even when a structure is defined intrinsically, our grasp of it and our ability to reason about it is usually grounded in our command of the informal ways to characterize it, including those that appeal to its symmetries. Similar points can be made with respect to Jacobs' other example, i.e., the characterization of Galilean spacetime in terms of the coordinate functions. This is a perfectly fine way to define that structure. Yes, it employs auxiliary tools, i.e. coordinates, but that does not, to my mind, make the characterization any less perspicuous.

Consequently, the structural differences between Maxwell Gravitation and Newton-Cartan Theory were understood well enough long before March's intrinsic definition of Maxwell

[41] Many thanks to James Read for mentioning this example to me.

Gravitation, and were the focus of much of the discussion. Specifically, the source of interpretive difficulty is that in Newton-Cartan Theory, one always has a connection, even if the matter distribution is insufficient to determine its value. This happens for instance in models with a single particle. The value of the connection in empty spacetime regions is not determined by the dynamics, so that NCT contains a multitude of non-isomorphic one-particle solutions (see e.g. Saunders [2013], March [forthcoming]). Saunders [2013] sees this as resulting from the introduction of surplus structure at an earlier stage, although this has been disputed (Wallace [2020], March [forthcoming]). In any case, we are looking at a fruitful and intricate debate which could and has been had in the absence of an intrinsic formulation of Maxwell Gravitation. This tells against a general requirement of intrinsicality. Indeed, one would actively hinder progress if one were to put debates such as this one on hold until an intrinsic formulation of the theories under discussion had been found.

We can furthermore draw on examples due to Wallace [2019] to show not only that intrinsicality does not invariably guarantee perspicuity, but even that non-intrinsic formulations can make certain interpretively relevant features of a theory especially clear. First, Wallace points out that one can define pre-relativistic spacetime theories in the setting of differential geometry, which amounts to defining their spacetimes as manifolds equipped with a connection. This should for all intents and purposes count as an intrinsic definition. Nonetheless, this definition is arguably less than fully perspicuous since it foregrounds features of the connection – whether it is flat or curved – as the distinguishing factor between pre-relativistic and relativistic physics, even though it is the difference between a global and a local geometry that truly marks the fault line between the theories. To make this salient however, it is preferable to define pre-relativistic theories over affine spaces rather than manifolds.

Moreover, as Wallace [2019] points out, in many cases the extrinsic definition of a structure is *more* intuitive and perspicuous than its intrinsic counterpart. This shouldn't be too surprising. After all, being presented with a structure via its symmetry group gives one a powerful tool for visualization. Thus, Wallace writes:

> To characterise Newtonian space-time – a conceptually fairly simple and intuitive spacetime – even in qualitative differential-geometric terms takes Friedman (1983, pp. 71-78) some eight pages. To do it properly requires familiarity with mathematics

> generally studied at graduate level. In contrast, it takes about a line to define the spacetime in terms of its symmetry group, and that line needs no mathematics beyond a little linear algebra.
>
> (Wallace [2019], p. 134)

Wallace further reminds us that "the 'coordinate-free' approach isn't as coordinate-free as all that." (Wallace [2019], p. 126) For instance, we use coordinates to define manifolds – that is just not evident because we simply write M at the end. Does Jacobs' intrinsicality condition amount to a demand that we characterize manifolds intrinsically? This is indeed possible, with the help of highly abstract and advanced mathematics (Wallace [2019], p. 131n5), but it is hard to see what epistemic benefits might accrue to that, or how a failure to do so should leave us with an opaque metaphysical picture.

To further strengthen the case against a requirement of intrinsicality, let me also point out how sweeping it is in ruling out certain common and indispensable ways of formulating theories. The paradigms of non-intrinsic definition for Jacobs are instances of mathematical "trickery" such as quotienting and coordinate-based characterizations of structures. But note that as defined, Jacobs' notion of extrinsic theories comprises many more theories than just those not defined exclusively in terms of geometric objects. Think about theories that employ configuration spaces, such as quantum mechanics or even certain formulations of classical mechanics. I take it that configuration space and the objects that live on it, such as the wave function, do not "explicitly represent the world's physical structure" – even though in the case of quantum mechanics this is somewhat controversial. (see Albert [2013], Ney [2021]) Such theories are therefore not intrinsic in Jacobs' sense. Phase space is another example of an indirect representation of physical structure. What is especially noteworthy here is that phase space carries structure of its own, i.e., symplectic structure. But this structure is not to be found in physical space. Rather, the structure of physical space is encoded indirectly in the phase space representation. We therefore have another instance of non-intrinsic formulation. A requirement of intrinsicality would have us do without any of these standard ways of constructing theories.

For a last example of how extreme the demand for intrinsicality is, take Euclidean geometry. There is a perfectly standard way of defining Euclidean geometry using a metric. This involves a mapping into the reals however, and so is not an intrinsic definition of Euclidean Geometry since it introduces an absolute standard of length. For an intrinsic

formulation, we must turn to Hilbert's admirable but rather bizarre axiomatization of Euclidean Geometry in terms of primitives such as a 'betweenness' predicate and a six-place (!) predicate for angle-congruence (see Field [1980]). I can't help but feel that the ordinary way of doing geometry is just fine. In sum then, a general requirement for intrinsicality seems much too strong since it would impede progress and would be cumbersome if not impossible to implement given the ubiquity of extrinsic methods. Moreover, we cannot derive a demand for intrinsicality from considerations in favour of perspicuous theories, since extrinsic formulation is not invariably an obstacle to perspicuity. The 'hidden flaws' argument draws on admittedly important cases in which extrinsicality really is an obstacle to perspicuity, the reformulation of electrodynamics in terms of equivalence classes [A] of vector potentials being a prime example. In this case, the superficial simplicity of the theory makes it all too tempting to overlook the inherent explanatory deficiencies of the theory. But since such cases are not universal, they cannot yield a general requirement of intrinsicality.

What is more, it seems to me that intrinsicality does not guarantee perspicuity. This further demotivates a requirement of intrinsicality of course, since if intrinsicality cannot be counted on to deliver perspicuity, there seems to be little reason left to insist on it. Let me therefore briefly conclude the discussion by explaining why I think intrinsicality and perspicuity are independent of each other. This means showing that we can have one without the other, and vice versa. I take it to be evident that there are theories that are simultaneously intrinsic and perspicuous, as well as theories that are simultaneously extrinsic and imperspicuous. This section has aimed at length to establish that there can be theories that are non-intrinsic but nonetheless perspicuous. To rehash but two examples, consider Earman's formulation of Maxwell spacetime and formulations of classical mechanics on phase space. I take the latter to be fully perspicuous because we understand completely how such formulations relate to equivalent formulations of classical mechanics on 3D Euclidean space, which in turn are perspicuous if any theories are. This leaves as the last case to consider theories that are intrinsic but imperspicuous. For a hypothetical example, suppose wavefunction realism were true.[42] Then any representation of quantum systems by a wavefunction on configuration space would be intrinsic. But this fact would be completely non-obvious, since wavefunction realism is widely rejected as an interpretation of quantum mechanics. Thus, if wavefunction realism

[42] For a defence of wavefunction realism, see e.g. Ney [2021]

were correct, then quantum mechanics would currently be a highly imperspicuous theory. Thus, the configuration space representation of quantum mechanical systems would be intrinsic but we would not be in a position to recognise that it is.

I conclude that there are no good reasons for insisting that reformulations must be intrinsic. A requirement of intrinsicality strikes me as overly demanding to an inordinate degree. Once one realizes the sheer amount of "non-intrinsicality" in the physical theories we use, Jacobs' vision becomes somewhat nebulous and open-ended. In particular, the 'hidden flaws' argument does not establish that there is a requirement of intrinsicality. What then does it show? The most natural suggestion by far is that it points towards a requirement of perspicuity: a reformulation of a theory can only warrant the judgment that the original theory's SRMs are equivalent if it is perspicuous in the sense that we possess a characterisation of its metaphysical picture of the world. Somewhat surprisingly, I believe that this also is too strong a requirement. Under certain circumstances, a reformulation can warrant our judgments of equivalence even before we have a perspicuous characterisation of its ontology and other metaphysical commitments. I will defend this claim in the next section.

5.3 Against perspicuity

To reject a requirement of intrinsicality is one thing, to reject a requirement of perspicuity is another. The idea that we are only licensed to judge symmetry-related models to be equivalent once we have a perspicuous characterisation of the ontology underpinning this equivalence is so central to standard forms of motivationalism (e.g. Moller-Nielsen [2017], Read and Moller-Nielsen [2020]) that some even doubt there could be a motivationalism that rejects it (Read, personal communication). I disagree: I believe there to be an intelligible form of motivationalism according to which judgments of equivalence for (non-isomorphic) SRMs of a theory must be warranted by a reformulation of the theory, and that this reformulation must be not necessarily perspicuous, but nonetheless recognisable as not deficient in explanatory and unificatory strength as compared to the original theory. Indeed, this is the motivationalism I defend. For this motivationalism not to collapse back into standard motivationalism however, it needs to be established that recognition of a reformulation's adequacy in the relevant sense can sometimes be gained even in the

absence of a fully perspicuous understanding of the reformulation's ontology. This is what I aim to do now.

In arguing against a requirement of perspicuity, I am positioning myself against the very strong intuition that only perspicuous theories allow us to assess their strengths and weaknesses in the way needed for comparing a theory to its reformulation. So, to begin to undermine this intuition at least somewhat, let me ask, what exactly does the motivationalist demand that in order to interpret symmetry-related models as equivalent, we need a "coherent metaphysical picture of the common ontology underpinning their equivalence" (Martens and Read [2020], p. 320) amount to? Once we probe into this question, we see that if it is to be plausible, it cannot amount to all that much.

First off, we can distinguish between a weak version of the demand and a strong one. The weak version has it that it suffices to present *some metaphysical picture or other* according to which the two models describe the same state of affairs. This picture will be given by some (perhaps partial) sophistication of the theory. The strong version on the other hand has it that we must identify the best candidate for the ontology underlying the models of the original theory, and that only then we are licensed to regard symmetry-related models as equivalent. To see the difference, consider two models of NGT related by a kinetic shift. If we impose the weak version of the demand for a shared ontology, it suffices to move to Galilean spacetime to consider the kinetic shift-related models equivalent. If we impose the strong 'shared ontology' demand however, we must identify the most plausible candidate for the ontology shared by the two models before we are licenced to treat them as equivalent. This would amount to constructing Maxwell Gravitation, I take it. And yet, we were clearly licenced to regard kinematic shift related models as equivalent already when Galilean spacetime became available, despite the fact that Gravitation Theory on Galilean spacetime faces its own problems – specifically, it exhibits a further symmetry, the so-called 'dynamic shift' symmetry. So, we see that motivationalists can at most insist on the weak version of the demand for a perspicuous account of the ontology shared by symmetry-related models. It is not the case that we need a "coherent metaphysical picture of *the* common ontology underpinning their equivalence" (Martens and Read [2020], p. 320, emphasis added). At best, we need some metaphysical picture or other that is preferable overall.

This already weakens the pull of the perspicuity condition: as the example suggests, what really justifies our judgment that the symmetry-related models in this case are equivalent – what does the heavy lifting, as it were – is the fact that we can find a reformulation that retains the original theory's explanatory strength while eliminating its symmetry-variant structure. It is not the fact that we have found a fully satisfactory ontology that could underpin the equivalence of NGT's SRMs – after all, the metaphysical picture offered by Gravitation Theory on Galilean spacetime is dubious in itself, since it is afflicted by further symmetries.

We can furthermore find cases in which a reformulation seems to warrant our judgment that certain models are equivalent despite the fact that the reformulation fails to be perspicuous. For a first example, take a representation of a quantum system by a wavefunction on configuration space. Its Fourier transform is another representation of the system. How can we know that they are equivalent? The usual argument has it that we can reformulate quantum mechanics as a theory set on an abstract Hilbert space, in which the configuration space representation and its Fourier transform appear as interchangeable representations in specific operator bases (see e.g. Butterfield [2018], de Haro [2021]). Crucially, none of the three representations – the configuration space representation, its Fourier transform, or the abstract Hilbert space representation – is perspicuous.[43] After all, there is much we do not yet understand about the metaphysical picture behind quantum mechanics. Nonetheless, the Hilbert space formulation clearly permits us to interpret the two original representations as equivalent. The reason for this is that we can recognise that nothing is lost in moving to the Hilbert space formulation, because the original representations can be recovered in that theory. There is no doubt than that the new theory is at least as good a theory as its predecessors, and therefore constitutes an adequate reformulation. We can reach this verdict despite the fact that the theory is not perspicuous.

A further powerful argument against a requirement of perspicuity can be derived from theories in which some ineliminable structure – ineliminable by our current mathematical tools at least – *must* be treated as encoding redundancy in order to get the theory to work. We find good examples of this in the treatment of the vector potential in various quantum

[43] One might think that nowadays, quantum mechanics is perspicuous since we have the Everett interpretation. Even so, there clearly was a time when quantum mechanics was completely imperspicuous. Even at that time, the Hilbert space representation allowed us to interpret the configuration space representation and its Fourier transform as equivalent, which is all the argument requires.

field theories such as quantum electrodynamics or quantum chromodynamics. Their formalism itself forces one to treat vector potentials related by a gauge transformation as equivalent, which is reflected in various mathematical corrections to the naïve way of incorporating the vector potential into these theories. For example, when quantizing electrodynamics to arrive at QED, care must be taken to accommodate the gauge redundancy since otherwise the path integral will be ill-defined and divergent. (Dougherty [forthcoming]) If one simply takes the configuration space of classical electrodynamics and the Maxwell action

$$S_M(A) = \int d^4x \frac{1}{2} A_\nu(\eta^{\mu\nu}\partial^2 - \partial^\mu\partial^\nu)A_\mu \tag{3.15}$$

as the basis for quantization, one ends up with an unsalvageable divergence in the path integral, for the reason that any vector potential contributes to the integral its entire family of gauge-related, physically equivalent potentials – see Dougherty ([forthcoming], pp. 18-20) for detailed discussion.

To tame the infinities that arise one must account for the redundancy in the vector potential. This can be achieved by means of various mechanisms. One way to do it is to fix a gauge (Healey [2007]). Another is to equip the underlying configuration space of the classical theory with additional structure which allows one to identify gauge-related potentials via parametrization – see once again Dougherty [forthcoming] for details. Dougherty calls this additional structure 'gauge structure' (Dougherty [forthcoming], p. 18).

The crucial point now is this: The fact that for the purposes of quantisation, the vector potential *must* be treated as encoding partial redundancy on pain of divergence compels us to interpret models of classical electromagnetism related by a gauge transformation as equivalent, independently of whether we possess a perspicuous reformulation of the classical theory. After all, the assumption that no redundancy attaches to the vector potential blows up the QED path integral, which is as good a reductio as one could hope to find. On either way of fixing the problem, i.e. gauge fixing or the introduction of gauge structure, we end up with a reformulated version of the classical theory which differs from the original theory in that it eliminates gauge freedom. And we can show that it is precisely this feature that makes the new theory amenable to unification with quantum theory when the old theory was not. Crucially however, the new classical theory is not straightforward to interpret and therefore imperspicuous. This is especially clear in the case in which the new theory employs gauge structure – doing so is a mathematical patch for dealing with

redundancy, but gauge structure does not seem to directly correspond to any physical structure, which makes the formalism metaphysically unilluminating. A general requirement of perspicuity is therefore ruled out by these examples – sometimes, we can know that a theory involves redundant structure even if we don't have a perspicuous reformulation that eliminates the surplus at hand.[44]

If my arguments so far have been correct, the 'hidden flaws' argument does not suffice to establish either the intrinsicality requirement or the perspicuity requirement. Nonetheless, it clearly points towards something important. What exactly is the lesson it teaches then? To my mind, it tells us that if a reformulation is to warrant our judgment that the original theory's SRMs are equivalent, we must be able to ascertain that the reformulated theory is adequate in the sense of being a better theory than its predecessor overall, as judged principally by the standards of explanatory and unificatory strength. To put a label on it, let us call this the 'recognisable adequacy' condition that a reformulation must meet.

At this point, the reader would no doubt appreciate a more fleshed out account of "recognisable adequacy". What features a theory might exhibit are necessary or sufficient for us to be able to judge whether the theory is adequate? I am afraid I cannot offer a full answer to this question – more research is needed to address it. As stated above, I believe that if a theory is perspicuous, we are in a position to assess its adequacy. Perspicuity is a sufficient condition in this sense. This section has aimed to show however that it is not strictly necessary. Nonetheless, I take it that the default route towards establishing a

[44] Indeed, one might think that the cases under discussion pose a problem not merely for the perspicuity requirement but for motivationalism more generally. The reasoning here would be roughly this: one way to tame the divergences that arise from the redundancy in the vector potential is to fix a gauge. To do so however is to move to a theory in which no surplus structure is removed – the resulting theory still involves a vector potential, although double counting is avoided by declaring all but one candidate vector potential unphysical. Nonetheless, the success of the resulting theory warrants our treating models of electromagnetism related by a gauge transformation as equivalent. In consequence, we are justified in interpreting these models as equivalent despite the fact that we have not constructed a theory that eliminates what is surplus in the vector potential, or so the argument might go. This argument against motivationalism strikes me as only superficially compelling. Gauge fixing is recognizably in the same family of solutions to problems with redundancy as taking the equivalence class of vector potentials related by a gauge transformation. This latter manoeuvre however is commonly recognized as a way of eliminating surplus structure, the objections discussed above notwithstanding. In essence, gauge fixing amounts to choosing a preferred representative from the equivalence class of gauge-related potentials, which in practice is what one ends up having to do even when one makes the equivalence class explicit – recall the dynamical equations for Maxwell Gravitation for instance. So, there is a very natural sense in which gauge fixing ought to count as a way of eliminating surplus structure after all, which means that the argument against motivationalism sketched here does not go through.

theory's adequacy goes via rendering the theory perspicuous through interpretation. Cases in which imperspicuous theories can warrant judgments of equivalence for SRMs seem relatively rare.

In conclusion, I have argued that while reformulation is crucial to judgments of equivalence for symmetry-related models, we do not in general need intrinsic reformulation, and not even a perspicuous metaphysical picture. More broadly, this chapter has argued that where reformulation and giving a perspicuous metaphysical picture of the theory's ontology come apart, it is the former, not the latter that is crucial to judgments of equivalence.

Given the importance the motivationalism defended here assigns to reformulation, the question arises how exactly we ought to go about reformulating theories exhibiting symmetries. The literature knows two strategies: reduction and sophistication (Dewar [2019]). In recent years, sophistication has been much more popular than reduction. In the next section, I will argue that reduction deserves a revival. As I aim to show, reduction is a much more fruitful and versatile strategy than is often assumed.

4 Reduction Revived

1 Introduction

As we saw, when a theory posits symmetry-variant structure, this is often considered a defect. In many cases, one has reason to suspect that this structure is surplus: that it has no role to play in an adequate description of physical reality.[45] This in turn suggests that the theory is in need of reformulation.[46] For some time, the dominant strategy for reformulation was *reduction*. (Caulton [2015], Dirac [1930], Nozick [2001]) Roughly speaking, to reduce a theory with symmetries is to reformulate it in such a way that the new theory is committed only to (some of) the symmetry-invariant quantities definable in the old theory. The symmetries of the original theory are thereby eliminated. Recently however, an alternative strategy called sophistication has found many advocates, even to the extent that its alleged superiority over reduction has almost become orthodoxy.[47] With sophistication, the aim is not to eliminate symmetries. Rather, one reformulates in such a way that in the new theory, any two symmetry-related models are isomorphic. Since in that case symmetries relate models that agree on all structure, it becomes arguable that none of the structure the theory posits is surplus – or that is the idea at least. An example that illustrates sophistication and reduction is that of electrodynamics formulated in terms of the vector potential. This theory exhibits a U(1) gauge symmetry, arguably indicative of surplus structure. The theory can be reformulated in terms of the Faraday tensor. This constitutes a reduction, since it eliminates the gauge symmetry. A sophistication of the theory on the other hand is obtained by rewriting it in terms of fibre bundles. Doing so renders models of the theory related by a gauge transformation isomorphic.

Its proponents argue that sophistication has several advantages. For one, it has been claimed that sophistications are easier to find than reductions (Dewar [2019]). Secondly, some maintain that in general, sophisticated theories have more explanatory power than their counterpart reductions (Dewar [2019], Jacobs [unpublished]). Relatedly, reductions

[45] On the link between symmetries and surplus structure, see for instance Ismael and van Fraassen [2003] and other essays in Brading and Castellani [2003]
[46] Moller-Nielsen [2017], Dewar [2019], [2024], Jacobs [2022-a]
[47] The label 'sophistication' is due to Dewar. Recent defenders include Moller-Nielsen [2017], Jacobs [unpublished], and Dewar [2019], although Moller-Nielsen's view is more subtle – he holds that isomorphic models may be interpreted as equivalent, but does not declare reduction an inferior approach to reformulation.

are often accused of positing "cosmic conspiracies" (see Arntzenius [2012], Jacobs [2024], Martens [2022]). The claim is that certain features of the symmetry-invariant quantities they deal in can only be demystified or explained if one treats the invariant quantities as relations between symmetry-variant quantities. For instance, that interparticle distances obey the triangle inequality is explained by the fact that they are distances between (symmetry-variant) locations in space (Jacobs [forthcoming]).

In light of these considerations, sophistication has been very popular. However, I will argue that the dethroning of reduction in favour of sophistication was premature. It seems to me partly the result of an unduly restrictive conception of reduction. This has hampered the search for reductions and deflected attention from success cases, as many possible ways to go about reducing theories were neglected. My aim in this paper is to demonstrate the resilience and flexibility of the reduction strategy by looking at a symmetry that *prima facie* seems very difficult to handle for proponents of reduction. That symmetry is the scale symmetry of Newtonian Gravitation Theory.[48] This example will illustrate nicely how a less restrictive concept of reduction gives rise to a more fruitful research programme. The resultant theory arguably has various advantages over its rival obtained from the sophistication strategy.

A further reason to be sceptical whether the sophistication strategy should be preferred to reduction is that the former arguably depends for its success on certain controversial doctrines; viz., anti-haecceitism and anti-quidditism.[49] Reduction on the other hand steers clear of such commitments. While this issue cannot be fully explored in this paper, I will outline the contours of the debate when comparing the two strategies below.

Section 2 presents the scale symmetry of NGT we will be focusing on. Section 3 then introduces the reduction and sophistication strategies as these are commonly conceived of in the literature. Section 4 presents a sophistication of NGT with respect to the scale symmetry. Section 5 investigates the project of reducing NGT with respect to the scale symmetry. I argue that on a restrictive conception of what reducing involves, the prospects of success are slim. Once certain restrictions are lifted however, a neat theory that deserves to be called a reduction of NGT with respect to the scale symmetry can be found.

[48] For recent discussions of scale symmetries see Dewar [2024], Sloan and Gryb [2021]
[49] Dewar [2019, 2024], Jacobs [unpublished], Moller-Nielsen [2017], and Read and Moller-Nielsen [2020] all appeal to these doctrines to defend their claim that isomorphic models are co-referential.

Section 6 compares the two theories. We argue that the reduced theory possesses various virtues that make it preferable to its sophisticated rival.

2 The scale symmetry

The scale symmetry is a symmetry of Newtonian Gravitation Theory. The models of NGT take the form

$$\langle D, E, T, x_i(t), m_i, \mathbb{R}^+ \rangle \tag{4.1}$$

where D is a set of particles, E is a three-dimensional Euclidean affine space, T is a one-dimensional Euclidean affine space representing time, the x_i are functions assigning positions in E at times t to the particles in D and the m_i assign values in the positive reals, representing masses, to each particle in D.[50] The DPMs of the theory are singled out by the following equations of motion

$$\ddot{\boldsymbol{x}}_i = \sum_{j \neq i} G \frac{m_j}{\left|\boldsymbol{r}_{ij}(t)\right|^3} \boldsymbol{r}_{ij}(t) \tag{4.2}$$

The scale symmetry of NGT is a dynamical symmetry which consists in a simultaneous scaling of the inter-particle distances by a factor λ, the particle masses by a factor μ, and the temporal distances between events by a factor τ, subject to the constraint $\lambda^3 = \mu\tau^2$. What is noteworthy about this symmetry is that it involves a simultaneous transformation of an internal quantity, mass, and external quantities, spatial and temporal distances.

3 Reduction and sophistication – a narrow account

The scale symmetry suggests that NGT should be reformulated. The strategies available to us are reduction and sophistication. But what exactly do these involve? In this section, we will give a restrictive, narrow definition. I am not claiming that everyone writing on this topic shares this narrow understanding of reduction and sophistication. Rather, the account I will present is supposed to capture a canonical set of ideas commonly associated with these notions, and, as far as I can tell, implicit in the way many think about them. I will subsequently argue that, at least in the case of reduction, we should widen our conception.

[50] Here, I follow the presentation in Jacobs [2022-b]

One characteristic of the narrow view is that reduction and sophistication are exclusively thought of as strategies to defuse symmetries *of a theory* – either by eliminating them, in the case of reduction, or by letting them relate only isomorphic models, in the case of sophistication. The first condition of the narrow account then is that theories R and S count as a reduction and a sophistication of a theory T, respectively, with regard to some symmetry h of T, if at all. This gives us Condition 1(a) (for reduction) and Condition 1(b) (for sophistication):

Condition 1: A theory can only be said to be a reduction or sophistication of a theory T with respect to some symmetry of T. If T is a theory with a symmetry h, then R(T) and S(T) are a reduction and sophistication of T with respect to h only if

(a) R(T) possesses no analogue of the symmetry h
(b) S(T) possesses an analogue of the symmetry h which relates only isomorphic models

Call Condition 1 the *'Symmetries of Theories' Condition*. One way to loosen it is to admit that we can reduce a theory also with respect to automorphisms of one or more of its value spaces, even if these automorphisms are not by themselves dynamical symmetries of the theory. For instance, even though mass scalings are not a symmetry of NGT, mass value space is plausibly invariant under scalings.[51] Consequently, we may want to reformulate NGT in terms of the invariants of these transformations, i.e. in terms of mass ratios. This would constitute a reduction of NGT not with respect to a symmetry of the theory, but with respect to a symmetry of one of its value spaces. We will see how this can be achieved in Section 5.

The second condition of the narrow account concerns the way a reduction R(T) of T with respect to a symmetry h must achieve the elimination of h. It states that R(T) must be formulated explicitly in terms of quantities that are definable in T and invariant under h. We can illustrate this with the example of the gauge symmetry found in the vector potential formulation of electrodynamics. If two vector potentials are related by a gauge symmetry, they will give rise to the same Faraday tensor. The Faraday tensor is therefore a symmetry-invariant quantity, and we can use it to construct a version of electrodynamics. According to the narrow account of reduction then, for R(T) to count as a reduction of T with respect

[51] We will see an argument for this in Section 4.

to h, any objects used in the construction of R(T) must be invariant under h. Call Condition 2 the *'Explicit Invariants' Condition*. We will loosen it later on, to allow theories to count as reductions even if they are presented with the aid of symmetry-variant quantities, as long as these quantities are mere gauge or quotiented out. Nothing more will remain of the 'Explicit Invariants' Condition than the demand that for a theory to count as a reduction of another theory with respect to some symmetry, it must eliminate that symmetry.

Commitment to something like the Explicit Invariants Condition is evident throughout Dewar's [2019] account of reduction. Thus, he presents reduction as the project of finding new dynamical equations, expressed in terms of symmetry-invariant quantities. A demand to formulate theories explicitly in terms of the quantities it treats as physical has also been defended by Jacobs [2022a], albeit in the context of sophistication. He calls it the demand for "intrinsic" formulation and argues that it makes the ontological commitments of the theory more perspicuous. While Jacobs is concerned only with sophistication, the demand for intrinsic formulation, explicitly in terms of physical quantities, is just as natural in the case of reduction, of course. I will argue however that the Explicit Invariants Condition is too strict and makes the search for reductions unnecessarily difficult. [52]

The third condition on reduction and sophistication concerns the question of empirical equivalence between T, R(T) and S(T). It states that any reduction or sophistication of a theory must be empirically equivalent to the original theory, and that this empirical equivalence is to be established via a process called *model recovery*. We say that T' recovers the models of T just in case there is a surjective mapping f of models of T onto models of T' such that for every M in T, f(M) is constructible from M and shares its empirical substructure.[53] Model recovery can be illustrated with the example of Trautman recovery. Trautman recovery is the way Newton-Cartan Theory recovers NGT, of which it is a reduction with respect to the so-called dynamical shift symmetry. A dynamical shift is a change in the absolute translational acceleration of the matter content of the universe. In NGT, dynamical shifts are constituted by a simultaneous transformation of the gravitational potential and the matter density, and are indeed symmetries of the theory. They can be eliminated by "geometrizing" gravity, which in particular involves excising the gravitational

[52] Interestingly, there is also a case to be made that a recovery of the old theory's dynamical equations in terms of symmetry-invariant quantities doesn't suffice for reduction. If this is true, it suggests that we should rethink the Explicit Invariants Condition more radically. This issue cannot be fully explored in this paper, but we will touch on it in Section 5.
[53] For the idea of an empirical substructure, see van Fraassen [1980]

potential. This transformation of the models of NGT yields Newton-Cartan Theory. For details, the reader is referred to Trautman ([1965]) and Malament ([2012]). The crucial point, in any case, is that any two models of NGT related by a dynamic shift induce the same model of NCT up to isomorphism, as Trautman's so-called recovery theorem shows.

The model recovery account of empirical equivalence is a particular version, appropriate to reductions and sophistications specifically, of a family of accounts of empirical equivalence which Martens calls "possibility checking" accounts. The basic idea behind possibility checking is that two theories are empirically equivalent just in case for every world w possible according to the first theory, there is a world w* possible according to the second theory which is empirically indistinguishable from w, and *vice versa*. One might also call this a 'veil of ignorance' account of empirical equivalence: the underlying thought is that even if we are ignorant of what world we are going to find ourselves in, as long as two theories are empirically equivalent, we can know that if the phenomena of the world we find once the veil is lifted are consistent with one of the theories, they will also be consistent with the other.

Call Condition 3 the *Empirical Equivalence Condition*. The Empirical Equivalence Condition is assumed, either explicitly or implicitly, in much of the literature on sophistication, reduction, and reformulation more generally. For example, Martens [2022], who sets out to find a mass comparativist alternative to Newtonian Gravitation, is explicitly limiting the candidate theories to those empirically equivalent to NGT.[54] [55] In consequence, he arrives at a theory that is mass comparativist, but absolutist with regard to spatial and temporal distances, since it is arguably impossible to construct a theory that exhibits full invariance under mass- and distance-scalings while remaining empirically equivalent to NGT. However, there is clearly something unsatisfactory about the resulting theory – a uniformly comparativist treatment of masses and distances would be preferable.

Elsewhere Dewar, a defender of the sophisticated absolutist approach to the scale symmetry, states that those who in light of this symmetry would instead rather be

[54] Although Martens relies on a criterion of empirical equivalence different from the model recovery criterion.
[55] For discussions of mass comparativism, see also Dasgupta [2013], Baker [unpublished-a, -b], Mundy [1987], Bigelow et al. [1988]

comparativists about mass and spatial and temporal distance[56] may want to "give up on trying to capture the full content of Newtonian gravitation, and seek to find some (*hopefully empirically equivalent*) alternative theory that admits arbitrary rescalings as symmetries, and hence is expressible purely in terms of these ratios" (Dewar [2024], p. 4, emphasis added) In this chapter, we will find an alternative theory that admits arbitrary rescalings as symmetries and is committed only to mass and distance ratios. But this theory will not be empirically equivalent to NGT.

Of course, settling the narrow account with the fairly demanding model recovery criterion of empirical equivalence makes it somewhat less encompassing of attitudes towards reduction than it could be. Niels Martens in his work on mass comparativist alternatives to NGT has articulated and tentatively defended a less exacting standard of empirical equivalence (Martens [2022]). I will discuss Martens' standard in an appendix to this chapter, but in any case, I am happy to grant that the narrow account could incorporate a more liberal criterion of empirical equivalence, since I will argue that we should consider certain theories reductions of other theories even though they are empirically inequivalent to their predecessors by any reasonable standard. If this is right, any account of reduction that insists on empirical equivalence in one form or another will be too narrow. So, nothing much hinges on spelling out empirical equivalence in terms of model recovery. I have included the model recovery account of empirical equivalence for the sake of definiteness, and also because it coheres with, and hence derives motivation from, the 'Explicit Invariants' Condition: if a reduction of a theory is supposed to be a version of the theory stripped bare of surplus structure, with only the symmetry-invariants remaining, it seems natural to demand that every model of the original theory should be retained in its pared down, surplus-free, form. Indeed, Dewar includes in his account of reduction the demand that "the new theory captures all the symmetry-invariant content of the old theory", ([2019], p. 493) and a version of the model recovery criterion is at work in Dewar's definition of reduction. (see esp. Dewar [2019], p. 495) So, Dewar insists quite explicitly on something like the model recovery criterion.

But demanding empirical equivalence, and model recovery in particular, strikes me as undermotivated. One motivating thought might be that a reduction should not sacrifice the

[56] Comparativists are of course one group who must pursue the reduction strategy, as the sophistication strategy yields commitment to absolute masses and distances – see Section 4, Jacobs [2022b], [Dewar 2024]

original theory's empirical success for the sake of eliminating structure. But that is consistent with looking for reductions that are predictively stronger than or approximately empirically equivalent to the original theory, perhaps even only under certain conditions – more on this in Section 5.3. A second line of thought, and one that motivates the model recovery criterion in particular, is the one intimated above, that a reduction should eliminate precisely the original theory's surplus structure while retaining all the physical structure. But this is artificially restrictive, since certain structural features might be suspicious and hence worth eschewing even if they are integral to the theory as originally formulated. For instance, as Bradley [2021] has shown, a standard of absolute rest is essential to Lorentz's ether theory. Nonetheless, it makes the theory unattractive, motivating reformulation. Similarly, relativistic Bohmian mechanics requires a preferred foliation which we would rather do without. Finally, the temporal metric and inertial structure of NGT, while essential to the theory, are regarded with suspicion by relationists, who seek to construct a theory free of these features (Pooley and Brown [2002], Barbour [2003]). So, there is no reason to limit the structure reduction is supposed to eliminate to surplus structure.

Below, I will argue in greater detail that the narrow account, as constituted by the 'Symmetries of Theories', 'Explicit Invariants', and Empirical Equivalence conditions, ought to be liberalised. Principally, the worry is that the narrow account does not encompass what look like clear success cases of reduction, either because those do not target symmetries of a theory, do not explicitly reformulate in terms of symmetry-invariant quantities, or fail to deliver empirical equivalence. Of course, this debate is partly definitional. However, one ought to examine whether the notions one defines get at anything important or whether they are artificially restrictive. Focussing on the narrow account of reduction can blind one to the manifold reasons one might have for reformulation, the range of mathematical tools available for that purpose, and the desiderata one should impose on a successful reformulation.

4 Sophistication with respect to scale

We now return to the scale symmetry. In this section, we will look at how the sophistication strategy deals with it. Putative sophistications of NGT with respect to the scale symmetry

have been put forth independently by Dewar [2024] and Jacobs [2022b]. Here, I will present Jacobs' theory, which I will call SNGT, for Scalable NGT.

The idea behind Dewar's and Jacobs' theories is that sophistication can be achieved through giving a more appropriate representation of NGT's value spaces. Take for instance mass value space. We have been representing it by the positive reals. But this arguably imputes too much structure to physical mass value space. The positive reals are rigid and have a preferred unit. We have reason to think that physical mass value space does not have these features. The fact that no choice of unit is privileged tells against the latter. The scale symmetry on the other hand indicates that physical mass value space is not a rigid structure. (Jacobs [2022-b]) To see this, we may follow Jacobs and appeal to a generalized version of Earman's [1989] symmetry principle SP1. In generalized form, it states (Jacobs [unpublished], p. 138):

(SP1) Any dynamical symmetry of T is a kinematical symmetry of T

Here, a *kinematical symmetry* is an automorphism of one or more of T's value spaces. SP1 presents an adequacy condition on our theories. (Earman [1989]) It encapsulates a demand to avoid surplus structure. For suppose that a theory violates SP1. Then there will be models related by a dynamical symmetry which nonetheless are not isomorphic because they involve structurally distinct assignments of quantities. This suggests that the value spaces in terms of which the theory's kinematics is defined are too highly structured.

It is easy to see that any theory which treats mass value space as a rigid structure will violate SP1, as scale transformations, while dynamical symmetries of the theory, will not be kinematical symmetries. A more appropriate representation of mass value space is therefore needed. This can be achieved by means of what is called an *additive extensive structure*.[57] Intuitively, an additive extensive structure is what one obtains if one takes the positive reals and "forgets", or "washes out", the multiplicative identity.[58] The resulting

[57] A precise mathematical characterisation of these structures can be found in Krantz [1971]. Wolff [2020, Ch. 5] also discusses these structures at length and in relation to the metaphysics of quantities.

[58] A perhaps even more illuminating characterisation of additive extensive structures describes them as totally ordered continuum-sized sets equipped with an associative addition function, subject to certain natural axioms. (see Jacobs [2022-b]) Obviously, when an additive extensive structure represents mass value space, the order compares masses and the addition function yields combined masses.

structure is no longer rigid. Its automorphism group is, precisely as desired, the group of uniform rescalings of the positive reals.

Considerations analogous to the case of mass show that the value spaces $\mathcal{V}_L$ and $\mathcal{V}_T$ for spatial and temporal distances should also be represented by additive extensive structures, albeit ones that include a 0-element acting as an additive identity. This is all we need to formulate SNGT. Its models are of the form

$$\langle D, W^3, T, \mathcal{V}_M, \mathcal{V}_L, \mathcal{V}_T, x_i(t), m_i, G\rangle \tag{4.3}$$

where

- D is a bare finite set of particles
- $\mathcal{V}_M$, $\mathcal{V}_L$, and $\mathcal{V}_T$ are additive extensive structures representing mass, length, and time interval value space, respectively
- W^3 (for Weylian space) is a three-dimensional affine space obtained by rendering Euclidean Space E dilation-invariant.[59] The vector space V_W associated with W^3 is equipped with a norm $||\cdot||: V_W \to \mathcal{V}_L$, mapping displacement vectors to points in length value space[60]
- T is a one-dimensional affine space such that the associated vector space V_T is equipped with a norm $||\cdot||: V_T \to \mathcal{V}_T$
- The x_i are particle trajectories through W^3; the m_i are assignments of mass
- $G: \mathcal{V}_M \times V_W \to (V_T \to V_W)$ is a function, replacing the gravitational constant of ordinary NGT. It yields an acceleration in accordance with the law of gravitational attraction

G is subject to certain constraints. (Jacobs [2022b]) These are that G be homogeneous of degree 1 in m, homogeneous of degree -2 in **r**, proportional to **r**, and that it satisfy

$$G(m, \mathbf{R}\boldsymbol{r}) = G(m, \boldsymbol{r}) \tag{4.4}$$

where **R** is a rotation matrix. The equations of motion are:

[59] I omit a precise mathematical definition of W^3 for the sake of brevity. In any case, it is clear enough what structure is meant. One intuitive way to obtain W^3 from E is to replace the inner product on V_E by the equivalence class of positive definite inner products on that space and to choose an appropriate norm
$||\cdot||: V_W \to \mathcal{V}_L$. This way, one eliminates the length scale defined by the inner product. In fact, this procedure involves some redundancy. It suffices to place certain restrictions on $||\cdot||$ to define W^3, but these are somewhat too involved to list here.

[60] In fact, Jacobs does not say exactly what structure is supposed to take the place of E in the models of SNGT. But W^3 is the obvious choice, as it is precisely Euclidean space made invariant under dilatations. So, I take it this is what he had in mind, and that we may specify the appropriate structure on his behalf.

$$\ddot{x}_i(t) = \sum_{j \neq i} G\left(m_j, \boldsymbol{r}_{ij}\right) \tag{4.5}$$

As required for a sophistication, the scale symmetry relates isomorphic models of SNGT.[61] But we observe the following: In this theory, mass value space is invariant under uniform scalings. W^3 is invariant under dilatations, in addition to the usual rotational and translational symmetries of space. Concomitantly, spatial distance value space is scale invariant, as is temporal distance value space. However, just like with NGT, these transformations, taken in isolation, are not symmetries of the theory. A mere scaling of distances for instance yields a non-isomorphic, empirically distinguishable model. In SNGT, this is ensured by the presence of G. G, here a function, not a constant, constitutes cross-value space structure (Jacobs [2022b]). It links the value spaces, partly rigidifying them in such a way that only compensatory simultaneous mass-distance-time scalings that obey the $\lambda^3 = \mu\tau^2$ constraint are symmetries of the theory. Note however that accordingly, not every internal symmetry of a value space is a symmetry of the theory. We will discuss whether this is a blemish of SNGT in Section 6.

5 Reduction with respect to the scale symmetry

5.1 The narrow account of reduction

As we saw, we can sophisticate NGT with respect to the scale symmetry. Can we also find a reduction of NGT with respect to the scale symmetry? If we adhere to the narrow account of reduction, then in all likelihood, the answer is negative. The 'Explicit Invariants' Condition tells us that any theory that is to count as a reduction of NGT with respect to the scale symmetry will have to be formulated in terms of the quantities invariant under this symmetry. What exactly are these quantities? *Prima facie*, it seems that we have the relative distances, both spatial and temporal, as well as the mass ratios.[62] The challenge is therefore to formulate a new theory directly in terms of those quantities, e.g. by providing a

[61] *Mutatis mutandis*, the proof of this is the same as Dewar's ([2024]) proof that in his sophistication of NGT, scale-symmetry related models are isomorphic.

[62] An important further fact that is preserved is whether the system has vanishing total angular momentum or not. This may be thought to be helpful for the task of finding a reduction. But since the angular momentum itself is not preserved, it is not straightforward to say what the corresponding invariant quantity is, in terms of which the reduction is supposed to be formulated.

Lagrangian or dynamical equations that involve only them. We must moreover ensure that the resulting theory satisfies the Empirical Equivalence Condition. But it is hard to see how to approach this task because any theory formulated in terms of those quantities alone would be invariant under arbitrary scalings of the masses and distances, whereas the dynamics of NGT is not – in NGT, varying the distance between two bodies for example influences whether they will orbit each other or escape from each other.[63]

Reduction with respect to the scale symmetry then poses a formidable challenge. To fully appreciate how difficult the task at hand is going to be within the confines of the narrow view of reduction, it helps to consider a somewhat different, but closely related project, which is to provide a relationalist alternative to NGT admissible by Leibnizian or Machian lights. Looking at this example in detail will prove instructive, as it will demonstrate both the limitations of the narrow view and a more fruitful way to think about reduction. Leibniz states his relationalist conception of space and time as follows:

> As for my own opinion, I have said more than once, that I hold space to be something merely relative, as time is; that I hold it to be the order of coexistences, as time is an order of successions. For space denotes, in terms of possibility, an order of things which exist at the same time.
>
> (Alexander [1956], pp. 25-6; cit. Pooley and Brown [2002])

When turned into a precise ontology, what is real according to this conception are the "order of coexistences [...] at the same time", i.e., (absolute) distances between particles in simultaneous configurations, and "the order of successions", i.e., (absolute) temporal distances between a configuration of particles at different times.[64] Thus, the Leibnizian conception countenances facts such as that two particles are currently two meters apart, that they were one meter apart one second ago, that their relative velocity is one meter per second, etc. Evidently, fewer matters are treated as factual according to this Leibnizian conception than are factual by Newtonian lights – such as the positions of particles relative to absolute space and time for example. In particular, no facts countenanced by the Leibnizian conception seem to distinguish a system of two spheres connected by a chord that is at rest from a rotating system of this kind, at least absent further bodies that could

[63] The relevance of this to the absolutism-comparativism debate was pointed out by Baker [unpublished-a, -b]. See also Martens [2022] for extensive discussion.

[64] I say 'absolute' instead of 'relative' since Leibnizians are usually not taken to be committed to scale invariance.

serve as reference points. This illustrates that the Leibnizian conception cannot distinguish between (global) inertial and non-inertial motion.

Consider a relationalist who subscribes to the Leibnizian conception. The task for them is to provide an alternative to the Newtonian theory that is committed to nothing but the quantities Leibnizians consider physical, i.e. absolute distances between particles and their rate of change with respect to time. Clearly, we can view this relationalist as engaged in a project of reduction. The extent to which this project can succeed has been investigated extensively in the literature.[65] We can draw a number of lessons for the narrow view of reduction from the results obtained.

The first concerns the 'Explicit Invariants' condition. Historically, a number of relationalist candidate theories were developed by giving a Lagrangian defined explicitly in terms of the distances between particles.[66] This is in line with the narrow account, which characterizes reduction as involving the formulation of equations directly in terms of the symmetry-invariants. But the theories obtained in this way failed to be empirically adequate – the issue was not so much that they were not exactly empirically equivalent to NGT, but rather that their predictions were blatantly wrong (Barbour [2003], Pooley [2013]). As Barbour [2003] explains, success for the relationalist programme only came when he and Bertotti abandoned this direct approach in favour of a more mathematically sophisticated one. We will look at the resulting theory, Barbour Bertotti Theory, in more detail below, as it will be the bedrock for our reduction of NGT with respect to the scale symmetry.

As an aside, let me point out that the 'Explicit Invariants' Condition is arguably also not *sufficient* for successful reduction, not even modulo satisfaction of the Empirical Equivalence Condition. To see this, note first that the Newtonian equations for an n-particle system of gravitating bodies can be reformulated explicitly in terms of the particle distances, as Lagrange and others building on his work showed (see Barbour [2001]). But this arguably does not suffice for successful reduction in this case. For the resulting equations have several unattractive features: first, they will introduce two new constants, denoting the total energy and the total angular momentum of the system. Secondly, they will contain the third derivative of the particle distances with respect to time, which means

[65] For a comprehensive survey, see Pooley [2013] and references therein.
[66] One such theory is due to Schrödinger (see Schrödinger [1925]).

that a solution will require that initial values for the second derivative of the particle distances be provided.

Several reactions to this result are possible. One possible conclusion is that Lagrange's reformulation can form the basis of a reduction of NGT to a theory only committed to structure acceptable by relationalist lights, albeit one that arguably possesses fewer theoretical virtues than NGT. After all, on account of the constants and third derivatives it involves, such a theory has very poor predictive and explanatory strength. More interestingly however, one might argue that Lagrange's reformulation does not at all support a theory that should count as relationalist.[67] The absolutist Newtonian theory arguably offers a more natural, illuminating account of the equations Lagrange obtained than any relationalist theory could. Therefore even an ostensibly relationist theory whose dynamical equations are Lagrange's equations should really be considered an absolutist theory "in disguise". This line of thought is reminiscent of Albert's (1996) and Arntzenius' (2012) arguments against configuration space realism in the context of Newtonian particle dynamics on the basis that the dynamics on configuration space is best understood as encoding a dynamics on three-dimensional space. It also echoes arguments due to Knox ([2011], [2014]) to the effect that certain theories, viz. teleparallel gravity and NGT, ought to be understood as positing not the geometrical structure in terms of which they are expressed, but instead the structure of General Relativity and Newton-Cartan Theory, respectively. However, while Knox considers cases in which theories arguably posit less structure than they wear on their sleeves, a relationist theory based on Lagrange's equations might be an example of a theory that is best understood as implicitly demanding more structure than is displayed. If so, it shows that reformulation in terms of invariants is not by itself sufficient for successful reduction. This in turn suggests that there is a tenuous connection at best between reduction and reformulation in terms of invariants, casting even more doubt on the Explicit Invariants Condition.

To return to the main discussion, we can also draw a lesson from the history of relationalism regarding the Empirical Equivalence Condition. One crucial insight that can be gained from Lagrange's recasting of the Newtonian equations in terms of inter-particle

[67] One reason to think this is that the theory violates Poincaré's criterion, which states that that the particle distances at some instant, together with their first derivatives with respect to time, should suffice as initial data to uniquely determine the evolution of the system. (see Pooley and Brown [2002]). It is often thought that a theory must satisfy Poincaré's criterion if it is to count as relationalist.

distances is that the equations simplify drastically when the total angular momentum of the system vanishes. In particular, the third derivatives with respect to time are then eliminated, and the relationally admissible data – distances and their first derivatives – suffice to predict the evolution of the system (Barbour [2001], Pooley and Brown [2002]). This suggests that the search for an elegant, predictively and explanatorily strong relationalist theory ought not to aim to recover all of NGT's models, but only the ones with vanishing overall angular momentum. In other words, we should want a relationalist theory that predicts that the total angular momentum of the universe vanishes. As we will see, Barbour Bertotti Theory does precisely that. Roughly speaking, the lesson is: if the total angular momentum of the universe does not vanish, we shouldn't want to be relationalists. But if it does, then we should want a relationalist theory that predicts this. Either way, there is no reason to look for a relationalist theory empirically equivalent to NGT. This of course undermines the Empirical Equivalence Condition.

Having examined the relationalist project, we can return to the scale symmetry. We note immediately that if the relationalist project was hard, finding a reduction with respect to the scale symmetry looks like it is going to be harder still. For in this case, we are working with nothing but mass and distance ratios. The relationalist at least had recourse to absolute masses and distances. If to reduce is to reformulate explicitly in terms of symmetry-invariant quantities as the 'Explicit Invariants' Condition tells us, and if the theory we are looking for must be empirically equivalent to NGT, then the prospects strike me as rather dim.

Moreover, I doubt that these prospects can be much improved even if one holds that in setting up the problem as that of constructing a theory in terms of mass and distance ratios we have suppressed invariants. To be sure, the imposition of the $\lambda^3 = \mu\tau^2$ side constraint means that the quantities invariant under the scale symmetry include not only the mass and distance ratios, but also quantities relating the absolute masses to the absolute distances. One might hope then that one can find a reduction empirically equivalent to NGT that is formulated not only in terms of mass and distance ratios but also in terms of these further invariants. Accordingly, Dewar suggests that we might try to work with "a quantity corresponding to the ratio between volume (length cubed) and the product of mass with the square of duration." (Dewar 2024, p. 4) Let us call this quantity η. We are then to use η in

constructing a reduction of NGT. But this approach – call it the *'mixed relations'* approach – faces major difficulties, both technical and interpretive.

The most immediate technical challenge is to formulate a dynamics featuring η. It is not at all clear how to go about this. There simply is no precedent for anything like it among the theories physicists have constructed so far. The definition of η in terms of mass and distances furthermore brings to light a number of interpretive issues for the 'mixed relations' approach. Were the approach to yield a candidate reduction, that theory's most immediately natural metaphysical interpretation would appeal to what Baker ([unpublished]) has called *comparativism with mixed relations*. Traditional, pure comparativism maintains that the most fundamental facts involving quantities are relational ones, such as perhaps

(C1) The earth is 83 times as heavy as the moon

or

(C2) Jupiter is five times as distant from the sun as the earth

or even

(C3) All things combined are x times as heavy as the earth[68]

These compare like with like, as it were – in this case, masses and distances, respectively. Comparativism with mixed relations on the other hand countenances at the fundamental level also 'mixed' relational facts, such as

(C-mx) The mass of this shelf in kg is twice its width in meters

(see Baker [unpublished]) Concomitantly, standard comparativism admits as fundamental quantities only pure mass and distance ratios etc., whereas mixed comparativism also allows for relations such as mass-(in kg)-to-distance-(in m) ratios or, in our case, η. But this gives rise to an interpretive problem: unlike pure relations such as mass or distance ratios, η and other mixed relations are *dimensionful*. η for instance has dimensions of $L^3M^{-1}T^{-2}$. But then it seems as though mixed relationalism fails to eliminate absolute masses and lengths.

[68] In light of Martens' [2022] theory of Machian Comparativism, which is formulated in terms of the ratios of the masses of bodies to the total mass of all matter in the universe, (see Section 5.2) it may be that these are the kind of facts comparativists should consider fundamental

The situation is in fact even worse: not only does it seem as though the mixed relations reintroduce the absolute quantities, they even appear to select a preferred unit for them. To see this, note that the facts mixed comparativism treats as fundamental, such as (C-mx), make ineliminable reference *both* to the quantities of (absolute) mass and length and to units in which these quantities are measured. This is so because the numerical ratio between the quantities will depend on a choice of unit – it would be not 2 but 2000 for instance if we measured the mass of the shelf in grams. As we saw above however, there are excellent reasons not to regard any one choice of unit as preferred. One may attempt to avoid the problem by including among the fundamental quantities the ratios of quantities in any system of units. But then the price for avoiding arbitrariness at the fundamental level is to introduce massive redundancy.[69] So, the prospects for the 'mixed relations' approach are dim.

These considerations may of course turn out not to be decisive. Nonetheless, they pose challenges for the 'mixed relations' approach that seem difficult to overcome. Many would see this merely as a vindication of the popular view that sophistication is the way to go and that reduction is not worth pursuing. This impression is reinforced by the fact that NGT can be sophisticated with respect to the scale symmetry, as we have seen. But this conclusion would be too quick. Rather, the lesson is that we should abandon the confines of the narrow view and rethink what a fruitful project of reduction might look like. This already became partly apparent when we examined the history of relationalism and the origins of Barbour Bertotti Theory, a theory the narrow account of reduction cannot accommodate. Thus, our discussion suggests that in thinking about reduction we should not be guided by the austere principles of the narrow account. Instead, we should learn from paradigmatic success cases such as Barbour Bertotti Theory. As we will see in more detail in the next section, BBT illustrates that virtually all of the conditions of the narrow account are overly restrictive and inhibit the search for reformulations of theories. The motivations for constructing a reduction, the mathematical tools involved, and the relation between the empirical content of a theory and that of its reduction can all deviate from what the narrow account demands. We therefore need a wider account of reduction. With that in hand, we will be able to tackle the scale symmetry without having to try to get the 'mixed relations' approach to work.

[69] See Sider [2020], p. 121 for a similar point

5.2 A more liberal account of reduction

To motivate and outline a liberalized alternative to the narrow account of reduction, we will look at two theories that ought to be considered reductions of NGT even though they violate certain conditions imposed by the narrow account. These two theories will not only illustrate ways in which the narrow account ought to be relaxed, but also help us find a reduction of NGT with respect to the scale symmetry.

The first example of a successful reduction of NGT is a mass comparativist version of NGT due to Niels Martens [2022]. We will see that it violates the 'Symmetries of Theories' condition. The 'Symmetries of Theories' condition demands that reduction must occur with respect to a symmetry of the theory. However, we can also reduce with respect to a symmetry of one or more of the theory's value spaces, even if those symmetries are not dynamical symmetries of the theory. Take for example mass value space. As we saw above, scalings are arguably a symmetry of this space. We may therefore want to look for an alternative to NGT formulated in terms of the invariants of those scalings. These invariants are the mass ratios. Martens ([2022]) shows that such a theory can be found. The theory he constructed is called Machian Comparativism (MC). Martens does not explicitly define the models of this theory, but we can do so on his behalf. We will take the models to be of the form

$$\langle D, E, T, x_i(t), \mu_i, (0,1] \rangle \tag{4.6}$$

where D is a set of particles, E is a three-dimensional Euclidian affine space, T is a one-dimensional Euclidian affine space, the x_i are functions assigning positions in E at times t to the particles in D and the μ_i assign values to the particles in D in the interval (0,1], representing the ratio of the mass of the i-th particle to the sum of the masses of all particles. Consequently, we require that the μ_i sum to 1. The dynamical equations for the theory are

$$\ddot{\boldsymbol{x}}_i = -\gamma \sum_{j \neq i} \frac{\mu_j}{\boldsymbol{r}_{ij}^2} \hat{\boldsymbol{e}}_{ij} \tag{4.7}$$

where γ is a constant.

One may wonder why Martens designs a theory that is mass comparativist but treats distances as absolute. He does so because he aims to construct a theory that is empirically

equivalent to NGT.[70] The result however is clearly a halfway house. Once we drop the Empirical Equivalence Condition, more natural theories will become available to us.

Consequently, our second example of a successful reduction of NGT that violates the narrow account is Barbour Bertotti Theory.[71] We will examine its mathematical formulation and its empirical content as compared to NGT. This will show that both the 'Explicit Invariants' condition and the Empirical Equivalence condition ought to be weakened. BBT comes in two versions, which I will label *Weak BBT* and *Strong BBT*, respectively. The theory was conceived to eliminate certain structural features of NGT, viz., its inertial structure and its absolute time. (Barbour [2003], Pooley and Brown [2002]) These features are not infrequently regarded as suspicious. As a consequence of the elimination, Weak BBT is invariant under time-dependent rotations and temporal distance scalings. Strong BBT furthermore eliminates the absolute length scale of NGT, and hence is additionally invariant under dilatations (Barbour [2003]). In their standard form, neither theory is invariant under mass scalings. Therefore, Strong BBT is not quite yet a reduction with respect to the scale symmetry. But this is easily fixed, as we will see in subsection 5.4. Thus, BBT will not only point us to a more fruitful way of thinking about reduction generally, but will also immediately yield the desired reduction in the specific case of NGT and its scale symmetry.

How is BBT obtained? Given an NGT system of n particles, we can define the configuration space of that system. But if two configurations are related by either a translation, a rotation, or some combination thereof, Weak BBT will treat them as representing the same physical state of affairs. One therefore quotients the configuration space by these symmetries to obtain the so-called *relative configuration space*. For Strong BBT, one quotients also by dilatations. This yields what is called *shape space*. A DPM of weak BBT (strong BBT) is then a curve image through relative configuration space (shape space) that meets certain conditions. Importantly, it is merely a curve image, not a curve. This captures the invariance of BBT under temporal scalings, and hence the elimination of absolute time.

[70] The reader may well wonder whether the theory is indeed empirically equivalent to NGT. If we spell out empirical equivalence in terms of model recovery, then clearly it is not. However, there is a robust sense in which MC is empirically equivalent to NGT nonetheless. Namely, note that every NGT model can be turned into a qualitatively identical MC model via the choice $\gamma = G\sum_i m_i$.

[71] For presentations of BBT, see Pooley [2013], Pooley and Brown [2002], and Barbour [1999, 2003]

Each DPM of BBT is a geodesic with respect to a certain metric. This metric is obtained from the standard, "kinetic" metric ds^2_{kin} on the Newtonian configuration space via a method called best matching. Best matching takes the Jacobi action on that space

$$I = 2\int dt\sqrt{T_{kin}(E-U)} = \int \sqrt{E-U}ds_{kin} \tag{4.8}$$

defined in terms of the kinetic metric and varies it along the symmetry transformations relating representations of the same physical configuration, i.e., rotations and translations. The length of the extremal path obtained in this way is the length of the geodesic in relative configuration space or shape space we were looking for. This method however only works if U meets certain conditions (see Pooley [2013]). These conditions in turn depend on whether one best matches with respect to only translations and rotations, as is the case for Weak BBT, or additionally also to dilatations, as with Strong BBT. With regard to U, Weak BBT requires that U be a function of the (absolute) distances between particles only. Strong BBT also requires that U be a function homogenous of degree -2 in the absolute distances.[72] Roughly speaking, this means that the potentials will be of the $1/r^2$ kind. Weak BBT furthermore requires that the total angular momentum of the system J=0; Strong BBT additionally requires E=0. We see then that either version of BBT makes stronger predictions than NGT – the three theories are not equivalent!

The lessons to be learned from this example are (1) that, contra the 'Explicit Invariants' Condition, reduction need not mean explicit reformulation in terms of invariant quantities and (2) that, contra the Empirical Equivalence Condition, a reduction of a theory T need not be empirically equivalent to T.

That BBT violates the Empirical Equivalence Condition is something we have already seen. That it also fails to meet the 'Explicit Invariants' Condition is illustrated by the fact that the definition of BBT helps itself very liberally to the full configuration space of NGT at various points. First, shape space is defined by quotienting configuration space. Secondly, the geodesics through shape space are obtained via variation of the Jacobi action on the full configuration space, since this is a mathematically more tractable problem (Barbour [2003], p. 1550). In consequence, the Lagrangians used in best matching, while scale-invariant, contain absolute distances and their changes with respect to absolute time as gauge quantities. The same is then also true of the equations of motion. Thus, it is quite

[72] This can be seen from Equation (9). Given that E=0 in Strong BBT, and T_{kin} scales quadratically in r, the Jacobi action is length scale invariant if and only if U scales inverse quadratically in r.

clear that BBT is not obtained by writing down equations explicitly in terms of relative distances. It certainly violates Jacobs' demands for "intrinsic" formulation. Nonetheless, BBT eliminates structure present in NGT, which provides a robust sense in which it constitutes a reduction.

In this section, I have argued that the tools for reduction should not be limited to the explicit use of symmetry-invariant quantities definable in the original theory. BBT makes use of quotienting and other tools, of which we should not deprive ourselves. Secondly, if we are willing to accept that BBT is a reduction of NGT, we should abandon the demand that a reduction be empirically equivalent to its predecessor.[73] One advantage of dropping the Empirical Equivalence condition is that when looking for reductions, we will no longer have to worry about identifying a set of invariant quantities sufficient for constructing a theory empirically equivalent to the theory being reduced. Call a set of quantities *complete* with respect to some theory T if it suffices to define a theory with a corresponding model for every equivalence class of symmetry-related models of the old theory T, and hence empirically equivalent to T. As Dewar [2019] points out, tensions may arise between the aim of finding a complete set of invariants and the aim of identifying a set of invariants that allow for a simple, elegant axiomatization of the reduced theory. This tension can be relieved by foregoing completeness and lifting the Empirical Equivalence condition. Of course, if the Empirical Equivalence condition is dropped, one will wonder what relationship ought to obtain between the empirical contents of a theory and its reduction instead. We will turn to this question in the next subsection.

5.3 Reduction and empirical equivalence

There must clearly be some relation between the empirical consequences of a theory and that of its reduction. If not equivalence, what exactly should this connection be? One suggestion is that if the two are not empirically equivalent, the reduced theory should make more predictions than the unreduced theory. To put this in terms of model recovery, the demand becomes something like the following:

[73] Note as further evidence here that electrodynamics in the Faraday tensor formulation only recovers a sector of the models of electrodynamics in the vector potential formulation

> If R(T) is a reduction of T, R(T) is a recovery of some subset of the models of T, i.e., every model of R(T) can be mapped to a suitable corresponding model of T, and the range of this mapping is a (possibly improper) subset of T.

This is what we find in the case of weak BBT. Weak BBT recovers precisely the models of NGT that have vanishing total angular momentum. It is also what we find in the case of the Faraday tensor version of electrodynamics, which recovers the models of the vector potential formulation that are based on a simply connected manifold. However, this proposal is not ultimately satisfying. The case of Strong BBT illustrates this. Strong BBT does not recover any of the models of NGT since, roughly speaking, models of BBT posit $1/r^2$ gravitational potentials, whereas the gravitational potential in models of NGT is of the $1/r$ kind. For this reason, no model of strong BBT is strictly empirically equivalent to any model of NGT, except for very limited cases such as zero- or one-particle solutions. This means that the attempt to characterize the relationship between a theory and its reduction in terms of partial recovery of the theory's models must be abandoned.

But I think that an alternative is available. To get some sense of this alternative, it is best to take a step back from the highly abstract account of empirical equivalence in terms of model recovery. Instead, we ought to look more concretely at the relationship that actually obtains between NGT and strong BBT, as this case presents a paradigm of a reduction that succeeds in accounting for the unreduced theory's empirical success. In essence, the achievement of strong BBT is to explain how the universe might "look Newtonian" without actually being Newtonian. As Barbour [2003] has shown, under certain circumstances, an island universe might look as though it was governed by potentials of the $1/r$ kind while in fact obeying potentials of the $1/r^2$ kind. One available explanation involves effective potentials – in essence, for systems meeting certain conditions, viz. being virialized and forming a globular cluster, one can find a potential of the $1/r^2$ kind that yields more or less exactly the same gravitational forces as the Newtonian gravitational potential (Barbour [2003]). Thus, if a universe meets certain conditions, a certain choice of potential allows BBT to reproduce the empirical predictions of NGT to an extremely high degree of accuracy. It is this relationship that we should want to obtain between the empirical contents of a theory and any theory that is to count as its reduction. Indeed, Barbour [2003] shows that there are more general ways to construct scale invariant potentials from Newtonian potentials. These scale invariant potentials yield as forces precisely the Newtonian forces plus a weak background cosmological force. Since the effects of the

cosmological force are minuscule relative to the effects of the Newtonian forces, we once again recover Newtonian behaviour to an extremely high degree of accuracy.

5.4 The scale symmetry revisited

With what we have said so far in hand, we can find a reduction of NGT with respect to the scale symmetry. Strong BBT is close to doing the trick. It can accommodate spatial and temporal scalings, but neglects the mass scaling symmetry. Luckily, fixing this problem is made simpler by the fact that in Newtonian Mechanics, masses are constants. As Barbour points out, "the dependence on the mass scale is not serious, because all masses are constant and can be expressed as dimensionless ratios of the total mass M." (Barbour [2003], p. 1560) To make this precise, we combine the basic ideas for BBT and for Martens' theory MC.

To begin with, it will be helpful to give a more general version of Machian Comparativism – call that generalization MC*. We must accommodate the $1/r^2$ potentials of Strong BBT. So, we formulate MC* in terms of potentials, and admit a suitably wide range of potentials over and above the mass comparativist version of the standard Newtonian potential

$$V_N = -\gamma \sum_{i<j} \frac{\mu_i \mu_j}{r_{ij}} \tag{4.9}$$

The models of the resulting theory MC* are of the form

$$\langle D, E, T, x_i(t), \mu_i, (0,1], U \rangle \tag{4.10}$$

where D, E, T, the x_i, and the μ_i are exactly as before, and U is a function representing the gravitational potential. The DPMs of the theory are singled out by the equation of motion

$$\mu_i \ddot{\boldsymbol{x}}_i = -\nabla_i U \tag{4.11}$$

where ∇_i is the vector derivative operator for the i-th particle.

How exactly we restrict the range of admissible potentials U won't matter much for our purposes, but let us say we demand that they be functions of the inter-particle distances only. This will still rule in $1/r^2$ potentials, such as

$$U_B = -\frac{V_N^2}{2} \tag{4.12}$$

Barbour [2003] points out that this potential is special in that for systems meeting certain conditions, it can replicate the predictions of NGT to a high degree of accuracy.

Having defined MC*, one can now proceed in exact analogy to Barbour and Bertotti in order to eliminate dependence on absolute distances and durations. In other words, one defines the Jacobi action for the configuration space of an MC* model and finds the geodesics on the associated shape space via best matching. The modifications of the Jacobi action induced by the move from NGT to MC* affect only the mass parameters and the constants occurring in the potentials – G becomes γ, etc.. Since the masses are constant, the conditions under which variation of the action yields a well-defined metric on shape space are not altered by the replacement of the masses by mass ratios. Consequently, the theory obtained is precisely a mass-scaling invariant version of BBT and hence, arguably, a reduction of NGT with respect to the scale symmetry. We will call this theory SBBT, for Scale-Invariant BBT.

6 Comparing SBBT and SNGT

I have argued that we can find a reduction of NGT with respect to the scale symmetry. But this really only helps rehabilitate reduction if the resulting theory is superior to, or at least not inferior to, the sophistication of NGT with respect to the scale symmetry, i.e., SNGT. Fortunately, this seems to be the case. SBBT possesses several virtues SNGT does not. First, SBBT is more predictively and explanatorily powerful than NGT, as SBBT predicts that the angular momentum and total energy of the universe are 0.[74] Secondly, SBBT is more structurally parsimonious than SNGT. Unlike SNGT, SBBT posits no cross-value space structure, and hence is distinguished by being fully invariant under mass scalings as well as spatial and temporal dilatations.

SBBT then is more parsimonious than SNGT in at least one respect. But there are many ways in which a theory can be more parsimonious than another. Martens [2022] introduces the notion of possible worlds parsimony. Roughly speaking, this is a measure of how many distinct metaphysical possibilities a theory admits. Introducing too many is often thought

[74] It is not entirely uncontroversial that this is a strength of the theory. For an argument that it is, see Pooley and Brown [2002]

pernicious, as it gives rise to underdetermination concerns. The classic example is the Leibniz shift: it is often maintained that Newtonian Gravitation Theory is committed to an infinite number of possible worlds which differ from the actual world merely by a static shift of all the matter around absolute space, i.e., by locating everything, say, three meters to the right of where it actually is.[75] SBBT however has the resources to collapse all these possibilities, since it only has a single model corresponding to each equivalence class of Leibniz shifted SNGT models. With respect to the theory pair under discussion then, we can ask whether SBBT is more metaphysically parsimonious than SNGT. Naively, the answer seems to be yes: for every one SBBT model, there is a whole equivalence class of Leibniz shifted SNGT models. More relevantly, given our focus on the scale symmetry, for every SBBT model, there is an equivalence class of mass-and-length scaled SNGT models. If different models denote different possibilities, then SBBT is more metaphysically parsimonious than SNGT.

The standard response by many proponents of sophisticated theories is to point to the fact that the symmetry-related models are in each case isomorphic. This means that the symmetry-related models do not posit any qualitative differences between the worlds they describe. At this point, proponents of sophistication usually invoke the doctrine of anti-haecceitism, which states that qualitatively identical worlds are numerically identical. This then allows them to argue that any two symmetry-related models correspond to the same possible world.[76] In the present case, this would render SNGT just as metaphysically parsimonious as SBBT.

But a few points should be noted. First, anti-haecceitism is a very controversial doctrine which faces many problems (see Fara [2009], Kment [2012]). Moreover, Dasgupta ([2011]) has argued that at the very least, a version of anti-haecceitism that can support substantivalism (and, by extension, quantity absolutism) is yet to be developed. Therefore, if we regard metaphysical parsimony as an important virtue, we should see value in constructing reductions, since, unlike sophisticated theories, their ability to deliver

[75] The *locus classicus* here is the Leibniz-Clarke correspondence (Alexander [1956]). For further discussion of Leibniz shifts and their implications, see *inter al.* Maudlin [2003, 2012], Arntzenius [2012]

[76] For anti-haecceitism as a doctrine undergirding sophistication, see Dewar [2019], Jacobs [unpublished]. For a general appeal to anti-haecceitism in the context of symmetries, see Hoefer [1996], Pooley [2006, 2013].

metaphysical parsimony is not hostage to the good standing of the anti-haecceitist doctrine.

There may of course be theory-specific reasons for regarding isomorphic models as representing the same possible world that do not depend on anti-haecceitism. But these tend to be stronger in the general relativistic setting than in the pre-relativistic setting. Roughly speaking, the reason for this is that in general relativity, the spatiotemporal, i.e., metrical, relations between spacetime points depend on the matter distribution in the universe, whereas in pre-relativistic physics they do not. This suggests that in NGT, one and the same spacetime could have carried a different matter distribution, while the analogue is in general not possible according to general relativity. Consequently, Maudlin ([1988], [2012]) has arrived at a view according to which Leibniz shifts relate distinct possibilities, whereas hole diffeomorphisms do not. Notably, Gordon Belot ([2018]) has argued that not only in NGT, but even in General Relativity, isomorphic models do not invariably denote the same possibility. The proponent of sophistication must either refute these views, which moreover put pressure on an indiscriminate attitude towards isomorphic models derived from anti-haecceitism, or accept that reduction yields a more parsimonious theory in the present case.

Yet another advantage of SBBT not shared by SNGT is that it possesses the following neat property: Every value space symmetry is a symmetry of the theory. Call this property *harmony*. SNGT is not harmonious. After all, mass value space is invariant under scalings, but mass scalings are not in general a symmetry of SNGT. Similarly, physical space as described by SNGT is invariant under dilatations. But once again, these are not symmetries of SNGT. Although I cannot defend this claim here, it strikes me that harmony is a desirable feature of theories. If this is right then SBBT has a decisive advantage over SNGT.

That being said, matters are not entirely straightforward. SNGT has some aces up its sleeve, too. Specifically, some will suspect that SNGT has certain explanatory advantages over SBBT for the reason that SBBT, but not SNGT, must postulate “cosmic conspiracies”. As Jacobs [2024] shows, reduced theories, specifically reduced theories that trade in symmetry-invariant relational quantities, frequently have to declare that these quantities conspire in certain inexplicable ways to behave exactly as though they were ontologically dependent on more fundamental absolute quantities when the theory assumes they are

not. For instance, (Leibniz shift invariant) inter-particle distances obey the triangle inequality, explicable by the fact that they are the distances between (Leibniz shift variant) points in space. It is supposed to be a great advantage of sophistications that unlike their counterpart reductions they admit the absolute quantities that explain the behaviour of the relational quantities.

Now, seeing as SBBT is a reduction, is it therefore a conspiracy theory? Crucially for our purposes, when a theory has to posit cosmic conspiracies, this will normally manifest in the definition of the theory's DPMs since the conspiratorial constraints will have to be imposed by hand. In this respect however, SBBT seems to be doing well. There are certainly no overt spatiotemporal conspiracies to be found in the formalism.[77] There is a highly visible aspect of SBBT however that reeks of conspiracy, viz. the constraint that the assignment of the μ_i be unitary, in other words, that $\sum_i \mu_i = 1$. At first glance, this might strike one as a glaring instance of having to put in by hand and without explanation a constraint that in the absolutist theory has the status of theorem – the basic pattern symptomatic of cosmic conspiracies.

But here, I believe defenders of SBBT has a good response available to them. It seems to me that there is a natural way to think about the assignment of μ_i on which the unitarity constraint turns out to be not an *ad hoc* imposition but a constitutive part of the definition of the assignment. Namely, one can (and should) think of this assignment as an assignment of relative weight. On this reading, it is perfectly natural, indeed required, to impose the unitarity constraint. After all, part of what it is to be an assignment of relative weight is that the total of the assignments should equal 1. So, unitarity can be thought of as not a conspiracy but a constitutive constraint. Consider the analogy with probability: suppose we have an indeterministic theory in which different outcomes get assigned probabilities. We would certainly demand unitarity for these assignments, just because that is part of what makes the relevant assignments assignments of probabilities. The same is true of assignments of relative weight. I would suggest therefore that the unitarity constraint is no more suspicious when imposed on the assignment of μ_i than it is in the case of probability.

[77] But see Arntzenius [2012], Ch. 1.5 for a careful critical discussion of Barbour's claim that best matching allows us to eliminate inertial structure, full engagement with which would go beyond the scope of this thesis.

To conclude, SBBT is an attractive theory which can claim many theoretical virtues in its favour. This alone suffices to establish that reduction is a project worth pursuing. Note also that, as the example of SBBT illustrates, reduction can yield new physics, which might be generalizable or more widely applicable. The discussion shows that it would be very premature to prefer sophistication to reduction *ab initio* and across the board. For any given theory, its reduction and sophistication will have to be compared carefully and on a case-by-case basis. Dismissing reductions out of hand is unwarranted.

7 But is it reduction?

I have defended the value of seeking reductions based on the example of SBBT. But does SBBT really deserve to be called a reduction of NGT?[78] SBBT eliminates the symmetry-variant structure of NGT at no cost in empirical adequacy. These are important constitutive features of reduction. Still, it is undeniable that the relationship between SBBT and NGT is much looser than the one we find in typical instances of reduction. The two theories are not strictly empirically equivalent and, what is worse, SBBT does not recover *any* of the models of NGT. Moreover, one could argue that with the 1/r^2 potentials, SBBT introduces structure that is not already definable in NGT. In what sense then should SBBT count as a reduction of NGT?

Let us consider the point that SBBT introduces new structure, i.e. 1/r^2 potentials. Is this an obstacle to regarding SBBT as a reduction of NGT? I would argue it is not, because there is a natural extension NGT* of NGT that admits a wider range of potentials – say, all potentials that are functions of the inter-particle distances only. We can view NGT as a restriction of NGT* to 1/r potentials. This viewpoint is well-motivated within the framework of Newtonian physics. After all, the specific form of the force law is not forced on us by the kinematics of the theory. Indeed, as Coffey (forthcoming) explains, part of Newton's project in the Principia was to provide a general framework for explaining natural phenomena in terms of forces of attraction and repulsion. The law of gravitation is thus to be seen as only one force among a potential multitude. So, there is a sense in which the 1/r^2 potentials were "in the air" all along. On Friedman's (2001) analysis of Newtonian Mechanics also, the

[78] I'd like to thank the audience at the Oxford Philosophy of Physics Seminar, particularly Samuel Fletcher and Eleanor March, for pressing me on this point. Much of what is to follow is derived from discussion with the audience.

gravitational law is located not on the constitutive side of the theory with the mathematical and mechanical principles but on the side of specific force laws to be determined empirically. In other words, nothing about the theory's constitutive core rules out 1/r^2 potentials. They can be accommodated within the Newtonian framework and so should be regarded as admissible. But even if this is considered an illegitimate move, it is not clear in any case what reasons there could be for insisting that a reduction with respect to structure variant under a specific symmetry should not introduce any new structure elsewhere.

Ultimately, whether or not SBBT should be regarded as a reduction of NGT is tied up with broader questions about theory interpretation and the role of reformulation, specifically reduction and sophistication, within interpretation. If the question one takes to be crucial to theory interpretation is what part of a theory's posited structure is dynamically efficacious and what part is inert and surplus, then SBBT is not a reformulation of NGT that is apt to answer core interpretive questions and hence should not count as a reduction. This question arises naturally when one combines a hermeneutic account of theory interpretation according to which to interpret a theory is to determine its exact truth conditions, with the view that a theory's truth conditions are to be discerned from a version of the theory free of surplus structure.

However, as I have argued in Chapter 2 of this thesis, the best way to think about theory interpretation for the purposes of the debate around motivationalism is not along hermeneutic but revolutionary lines. The crucial question for revolutionary interpretation is: can a theory's perceived flaws (such as for instance surplus structure) be remedied through reformulation? Nothing about the project of revolutionary interpretation rules out that this could involve quite radical reformulation. If this way of thinking about theory interpretation is on the right track, then the appropriate standards for successful reduction are merely that the reduced theory (i) eliminate the old theory's symmetry-variant structure and (ii) match or exceed the old theory in predictive and explanatory success. SBBT constitutes a reduction of NGT in this sense. Indeed, SBBT not only satisfies criteria (i) and (ii), it can also *explain* the empirical success of NGT. In other words, from the point of view of SBBT, the success of NGT is not a miracle or a coincidence. On what I consider the most fruitful way of thinking about theory interpretation and the role of reformulation within it therefore, SBBT is very naturally labelled a reduction of NGT.

Appendix. Martens on Empirical Equivalence

Are NGT and MC empirically equivalent? Martens [2022] says that they are, and I agree. But in virtue of what are they empirically equivalent? Here, Martens and I disagree. I will criticize Martens' argument for their equivalence and then present my own. I hope that in this way, we can obtain a clearer sense of what checking for empirical equivalence requires.

First, an important observation. As Martens notes, MC is not equivalent to NGT according to the model recovery criterion. To see this, consider a Newtonian world containing nothing but two planets of masses m_1 and m_2, respectively, traveling in opposite directions, separated by a distance r_0 at time t_0 and with initial relative velocity v_0. Depending on the initial values, the planets will either collide at some point, or they will travel further and further away from each other. Suppose we choose the initial values such that the planets escape each other. Then, holding r_0 and v_0 fixed, there will be some factor μ such that scaling m_1 and m_2 by μ turns this escape solution of NGT into a collision solution. This however means that there are empirically distinct models of NGT that agree on r_0, v_0 and the ratio of the planetary masses, i.e., on all the initial data available in a model of MC. Therefore, there will be a unique model of MC in which the planets are at distance r_0 at t_0, with initial velocity v_0 and mass ratio m_1/m_2. Depending on the value of γ, this will either be an escape solution or a collision solution. Either way, one of the two NGT models we have considered cannot be matched to an appropriate model of MC. MC therefore fails to be empirically equivalent to NGT according to the model recovery criterion.

However, Martens argues that on a more appropriate account of empirical equivalence, MC will turn out to be empirically equivalent to NGT. The account he proposes is still within the paradigm of the possibility checking, or 'veil of ignorance' approach to empirical equivalence. It merely aims for a more realistic account of when two worlds are empirically indistinguishable. For what are the empirically accessible data in a world? Martens considers two answers. According to the first, what is observable are not absolute distances, spatial or temporal, but merely distance ratios, as well as angles. This suggests that two worlds whose particle trajectories are similar, i.e. agree on angles and relative distances, are empirically indistinguishable. Call this the *Shape Criterion* of empirical equivalence. It gives us more leeway for possibility matching, since in the case at hand we no longer have to insist that the matched models agree on r_0 and v_0. If similarity of

trajectories suffices, we can avail ourselves of scaled solutions. The second answer Martens considers is more radical. It states that only topological facts, such as whether particle trajectories intersect, are empirically accessible. While this gives us even more leeway for possibility matching, it is arguably too lax a standard of empirical equivalence. In any case, the criticisms I have of the Shape Criterion generalize to the Topological Criterion, and so we can focus on the former.

Let us return then to the pair of Newtonian worlds. They agree, recall, on the r_0 and v_0 as well as the mass ratio m_1/m_2. But the absolute masses are greater in the second world, so that it is a collision world whereas the first world is an escape world. Suppose, without loss of generality, that the corresponding comparativist world, which agrees on r_0, v_0 and m_1/m_2 is also a collision world.[79] Then the Newtonian escape world is left without a match. However, we can find a Machian match as long as we insist merely on similarity, not congruence. Scaling r_0 and v_0 by a factor of

$$\sqrt[3]{\gamma / G(m_1 + m_2)} \tag{4A. 1}$$

will turn the Machian collision solution into an escape solution similar to the Newtonian one. More generally, given a Newtonian world, scaling the distances and velocities by a factor of

$$\sqrt[3]{\gamma / G \sum_k m_k} \tag{4A. 2}$$

yields a similar Machian world. The upshot is that if we treat similarity, i.e. agreement on shape, as the appropriate standard of empirical indistinguishability underlying the possibility checking approach, MC turns out to be empirically equivalent to NGT.

I have two worries about this argument. The first is that its plausibility seems to hinge on ignoring the role idealization plays in the way we use theories like NGT and MC to represent the world. After all, the idea that a world in which two objects are one meter apart is empirically indistinguishable from a world in which two objects are two meters apart is only convincing if we think of the objects involved as point masses. But this is of course merely an assumption we make in modelling Newtonian systems. Once we think of the bodies involved as spatially extended, it no longer seems plausible that worlds differing by a distance scaling are empirically indistinguishable. Perhaps one ought to think that

[79] This will depend on the value of γ of course, but what follows will hold *mutatis mutandis* if it is an escape world.

extended objects in a Newtonian world will be scaled alongside the velocities and inter-particle distances in the corresponding Machian world. But that does not seem right – they ought to be as they actually are, except for the properties explicitly varied across the worlds.

In any case, even if Martens were right that the Shape Criterion is sufficient for empirical equivalence, it seems to me that once we get clearer on the roles G and γ play in NGT and MC, respectively, a much neater and clearer account of their empirical equivalence becomes available. The need for Martens' complicated argument only arises from a confused understanding of the status of G and γ respectively in the two theories. Martens believes that his involved account of the empirical equivalence of NGT and MC is our only resort due to the fact that according to NGT and MC, G and γ each take some fixed value that is constant across all possible worlds.[80] After all, the task of finding a Machian world empirically indistinguishable from our Newtonian escape world would be trivial if we were allowed to adjust the value of γ. In that case, we could lower the value appropriately in order to obtain an escape solution to the dynamical equations of MC, and wouldn't have to resort to complicated arguments regarding the indistinguishability of length-scaled solutions. Scaling G or γ however, Martens claims, goes against "the rules of the game".

But even if we grant that, the exact value these constants take can only be determined through experiment – in that sense, G and γ are more akin to parameters that must be fitted to the data. After all, when Newton found NGT, he did not assign any particular value to G. Indeed, how could he have? Instead, determining that value was a matter of empirical investigation. To return to the veil of ignorance idea, while G and γ feature as parameters in the theories we are comparing on this side of the veil, the exact values of G and γ lie beyond the veil and must be determined a posteriori. In consequence, even if NGT and MC require the value of G and γ to be constant across all possible worlds, in any particular possible world the observational evidence will determine values for G and γ that allow NGT and MC to account for the phenomena in that possible world.

It seems then that the salient questions we ought to be asking when we try to establish whether NGT and MC are empirically equivalent are the following:

[80] This is especially clear in Section 6.1 of Martens [2022]

- Suppose that when the veil of ignorance is lifted I find myself in a Newtonian world. Is there some assignment of a value to γ such that there is a model of MC that adequately describes the phenomena in this world under that assignment?
- Suppose that when the veil of ignorance is lifted I find myself in a Machian world. Is there some assignment of a value to G such that there is a model of NGT that adequately describes the phenomena in this world under that assignment?

The answer to both questions is of course yes, and they seem to be the questions that capture the pre-theoretical intuition that NGT and MC are empirically equivalent much better than Martens' account.

5 Motivationalism about physical possibility

1 Introduction

In this chapter, I want to use the example of NGT and SBBT to point out a natural extension of the motivationalist viewpoint. The basic pattern of motivationalist thought is the following: our theories have various apparent shortcomings. These shortcomings motivate us to reformulate our theories so as to improve them. However, in the absence of a better theory, we cannot be certain that the features of our theories we deem problematic are dispensable, nor can we be sure that they do not accurately represent how things are in the world after all. So far, when discussing such problematic features, we focussed on symmetry-variant structure. But theories can be flawed also in other ways. For instance, a theory may possess strange models – models that we judge not to represent physical possibilities. While examples of this are bound to be controversial, we frequently find models in our theories with strange and arguably pathological features – take singularities for instance, or Norton's dome, or Gödel's solution to the Einstein field equations. Since we might have strong reasons to believe that these models do not represent ways things might have been, they can motivate reformulation just as much as symmetry-variant structure can. The aim in this case is to eliminate these models from the theory. For example, in the case of NGT and SBBT we see the elimination of a wide range of models: any model of NGT in which either $L \neq 0$ or $E \neq 0$ is not recovered in SBBT. But reformulation need not always aim at eliminating models. In other cases, it might be the fact that a theory *doesn't* admit certain models that motivates reformulation. Below, we will see that certain versions of NGT do not allow us to consistently describe universes with infinitely many homogeneously distributed massive bodies. This is a flaw that must be remedied by moving to mathematically more sophisticated formulations of the theory.

The reason it matters what models our theories admit is of course that these models guide our beliefs about what the world could be like or could have been like. Our reasoning here is governed very roughly by the principle that whatever is consistent with our best theories is nomically possible (and indeed perhaps the case!) whereas everything that is inconsistent with our best theories is not nomically possible.[81] Thus, for instance, when it was first shown that General Relativity predicts black holes, some people began to believe

[81] In this chapter, I use 'physical possibility' and 'nomic possibility' interchangeably.

that such objects, however strange they may be, could be out there in the universe. The question for our purposes is whether they were right to do so, or whether those who initially refused to believe in the possibility of black holes had the correct attitude. As in the case of motivationalism about symmetry-variant structure, motivationalism about strange models is at heart a thesis about how we ought to form beliefs about the world on the basis of our best theories. With symmetry-variant structure, the central question is: given that this structure plays a role in my theory, should I believe that it is physical? With strange models, the question is: given that these models occur in my theory, should I believe that they represent a genuinely possible way things might have been? Broadly, the motivationalist answer will of course be that we should think of every model as representing a physical possibility until we have a reformulation that eliminates the model or at least a diagnosis of why the model is pathological but nonetheless arises in the theory. Let us call this view motivationalism about nomic possibility.

Since nomic possibility and its epistemology fall outside the scope of the thesis, which is concerned with the epistemology of equivalence, I will not aim to undertake a full examination of this topic. In consequence, this chapter is less systematic and less thorough than the other chapters. Nonetheless, I hope to at least put a novel form of motivationalism on the table and to explore its contours to some extent. I also want to register some initial sympathy for motivationalism about nomic possibility. It can be seen from real historical examples that one does well not to rule out the physicality of any of a theory's models without a principled reason. As Christian Wüthrich (personal correspondence) has pointed out to me, in the early days of general relativity, black hole solutions to the Einstein field equations were often dismissed as unphysical by physicists. The discovery that they describe genuine physical objects was in turn one of the most impressive confirmations of any theory in the history of physics.

This chapter only has three sections in total. Section 2 goes into some detail on the relationship between models and modality – between what models a theory possesses and what is possible according to the theory. Section 3 then examines some examples of reformulation that either eliminates or adds models. On the basis of these examples, some general desiderata a motivationalist about nomic possibility might have for successful reformulation are articulated.

2 Models, Modality, Motivation

The examples of the previous sections all are models that we have reason to believe do not correspond to genuine ways things might have been, despite the fact that they occur in generally well-confirmed theories. They therefore undercut at least to some extent the status of these theories as straightforward guides to what is nomically possible. But what exactly is the general link between our theories' models and nomic possibility supposed to be in the first place?

It is a commonplace that scientific theories concern not only how things actually are but also how they might have been. The modal plays an indispensable and varied role in physics. To take an example from Baron, Le Bihan and Read [2025], Boltzmannian statistical mechanics defines the entropy of a macrostate in terms of the number of possible microstates which would realize it. Secondly, many physical theories are formulated on what is called a phase space. But a phase space is the collection of all possible states of the system (see Williamson [2017]). Furthermore, considerations of unrealized possibilities play a crucial role in scientific explanation. Variational principles such as the principle of least action explain a system's behaviour in terms of the fact that the realized behaviour is distinguished among the alternatives, for instance by minimizing the action. Mathematically, calculations of, say, a particle's path are performed by varying the action along possible alternative paths. As Baron, Le Bihan and Read [2025] point out, such modal considerations are also crucial to other forms of explanation, most prominently equilibrium explanations. In general form, an equilibrium explanation explains the state of a system by showing that any other state is unstable and will evolve into the state that is actually realized.

In consequence, we routinely turn to our theories to learn what had to be, what might have been, and what could not have been. Maudlin for instance writes:

> A physicist who accepts Einstein's General Theory of Relativity will also believe that it is physically possible for a universe to be closed (to collapse in a Big Crunch) and possible for a universe to be open (continue expanding forever). This is especially evident since we don't yet know whether our own universe is open or closed, so empirical data are still needed to determine which possibility obtains. But even if the issue were settled, the laws of gravitation, as we understand them, admit both possibilities.
>
> (Maudlin [2007], p. 7)

For a different example, it has been argued that spacetime is contingent, i.e., that it might have been that there is no spacetime, on the basis that only some but not all models of quantum gravity give rise to an emergent spacetime.[82]

In the background of these arguments is a general principle, according to which a theory's models represent what is possible according to the theory. Van Fraassen for instance states

> You can think of the models [of a theory] as representing the possible worlds allowed by the theory; one of these possible worlds is meant to be the real one. To believe the theory is to believe that exactly one of its models correctly represents the world (not just to some extent, but in all respects).
>
> (van Fraassen [1980], p. 47)

Jacobs [2023] has proposed that we make this idea precise by introducing two principles, representational soundness and completeness:

> **Representational Soundness**. For any reasonable representational convention C: if M is a model of theory T, then M represents a world W that is physically possible according to T under C
>
> **Representational Completeness**. For any reasonable representational convention C: if W is a world that is physically possible according to T, then there is a model M of T that represents W under C
>
> (Jacobs [2023], p. 121)

These principles are of course not entirely uncontroversial. Butterfield's [1989] account of the hole argument for instance denies soundness. Completeness is perhaps also a problematic principle when applied to effective theories, i.e., theories that explicitly incorporate a domain restriction.[83] This suggests that the principles might need some refinement. We will bracket these concerns here and assume that at least suitably qualified versions of Jacobs' principles are correct.

Jacobs' principles support two patterns of inference, assuming we believe a given theory to be a reliable guide to nomic possibility: on the basis of Representational Soundness, we can infer physical possibility from consistency with the theory. Call such inferences possibility inferences. With Representational Completeness, we can infer physical

[82] Christian Wüthrich suggested this to me in personal correspondence. See also Huggett and Wüthrich [2017] for discussion of the contingent emergence of spacetime in quantum gravity.
[83] See Rivat [2021] for detailed discussion of how best to understand the notion of an effective theory.

impossibility from inconsistency with the theory. Call such inferences impossibility inferences.

Of course, not just any theory will support reliable modal and counterfactual inferences – if we trusted all models of NGT for instance, not just the ones within its domain of applicability, we would conclude that objects can achieve superluminal velocities. There are difficult questions to be answered therefore concerning which of our theories are apt to tell us what is nomically possible and impossible. For a discussion of such questions see inter alia Baron, Le Bihan and Read [2025]. We will not address these questions in detail. For our purposes, it will suffice to note that given their broad background commitment to scientific realism, motivationalists will treat at least some theories as a guide to nomic possibility. More specifically, a motivationalist attitude towards at least our best theories will take roughly the following form: Our initial beliefs ought to be that whatever is consistent with the theory is possible, and whatever is inconsistent with the theory is impossible. Baron, Le Bihan and Read express an attitude of this kind when they state that "the models of an effective theory are innocent until proven guilty: our default position is to assume that they all approximate physical possibilities." (Baron, Le Bihan & Read [2025], p. 10) Should the verdicts derived from a theory in this way strike us as doubtful, for instance because the theory admits solutions we do not deem nomically possible, we are motivated to reformulate the theory in a way that excludes these solutions. Unless we find a reformulation that achieves this however, the initial verdict stands.

Examples of strange models that might strike us as unphysical abound. Earman [1995] for instance has argued that certain models of General Relativity should not be thought of as representing physical possibilities on the grounds that they contain closed timelike curves and therefore give rise to paradoxes of time travel. Any solution of GR that does not admit a global time orientation might seem unphysical. The paradoxes of time travel and retrocausality have also led people to argue against the possibility of tachyons. As is known, there exist tachyonic irreps of the Poincaré group.[84] However, Pirani [1970] for instance has argued that in the context of special relativity, admitting tachyons leads to paradox.[85]

[84] Many thanks to Adam Caulton for pointing me towards this example.
[85] But see Arntzenius [1990] for an argument to the effect that once proper equations of motion for tachyons are constructed, paradox is avoided.

In non-relativistic physics also, we find models that do not look as though they represent genuine possibilities. A good example here is Norton's Dome, specifically its non-stationary solutions. In the context of Newtonian physics, Norton's Dome is a system in which a point mass rests on top of a dome of a certain shape in a gravitational field (Norton [2003]). A number of distinct time evolutions of the system are compatible with Newton's laws: the point mass may remain at rest, or it may begin rolling off the dome at an arbitrary point in time and in an arbitrary direction. One might think that the rest solution is privileged, and that it alone corresponds to a genuine physical possibility. For a different example, take the framework of Newtonian mechanics that treats forces as primitive. In this framework, non-conservative forces are admissible, even though we might think that they are not real, at least not at a fundamental level. One might therefore prefer a potential-based framework of Newtonian mechanics, in which it can be derived that all forces are conservative.

These examples highlight possible breakdowns of possibility inferences. But impossibility inferences, too, can strike us as bad. Most often, it is certain possibility inferences that are seen as unreliable: some models of our theories are not to be trusted. But one should note that impossibility inferences, too, can be called into question. We might suspect for instance that a theory excludes certain possibilities simply because the mathematics its formalism relies on is not adequate for modelling them. This might be overcome by casting the theory in a framework that employs more advanced mathematical tools. We will see an example of this in the next section.

3 Paradigms of Reformulation

The aim of this section is to determine what reformulation should ideally look like from a motivationalist perspective. On the basis of examples, we can isolate some general characteristics of successful reformulation. As we have been noting, in most cases the aim of reformulation is to eliminate models that strike us as pathological in one way or another. Consequently, section 3.1 looks at these cases in detail. However, in some cases reformulation can also aim at adding new models. We look at this possibility in Section 3.2.

3.1 Eliminative Reformulation

One difficulty motivationalists have to contend with when formulating criteria for adequate eliminative reformulation is that it might seem trivial to exclude certain models: after all, when we find models that strike us as strange, why should we not simply declare that our new theory includes all the models we like and excludes all the models we don't like? For their view to have some substance then, motivationalists must tell us what conditions a reformulation has to meet if it is to warrant our belief that certain models do not represent physical possibilities.

One way to eliminate models of a theory is to add new axioms. In cases in which elimination is achieved by these means, motivationalists can content themselves with demanding that the new axioms ought to satisfy general norms of good theorizing: they should not be ad hoc, they should be independently motivated, general, "law-like", simple, elegant, etc. Of course, what exactly the super-empirical criteria are by which a theory's axioms ought to be judged remains controversial, but that is a more general debate, outside of the special concerns around motivationalism.

But not every case in which models are eliminated by reformulation is one of simply adding new axioms. The example of SBBT and NGT illustrates this very nicely. As we noted in the previous chapter, SBBT only recovers a subset of the models of NGT. For one, models with non-zero angular momentum are eliminated. Besides that, models of NGT with non-zero total energy cannot have a counterpart in SBBT. This is significant because it means that "pure inertial motion, for which $U = constant$ and $E > 0$, is not allowed" (Barbour [2003], p. 1546). Barbour believes it to be a flaw of NGT that it admits such pure inertial motion. As he explains, this is a consequence of the fact that NGT illegitimately divorces potential and kinetic energy, treating them as independently variable. In any correct treatment however, "[t]he two forms of energy must come as inseparable twins and their sum must be zero" (Barbour [2003], pp. 1546-7).

It is worth highlighting a number of key features of this example. First, the reformulation eliminates a class of models – models with $L \neq 0$ or $E \neq 0$ – in a principled, non-ad hoc way. It is not as though one simply points at certain models and manually excludes them from the theory without a deeper account of what warrants their elimination. Rather, the new

theory is such that the eliminated models *could not* have been recovered – the structure the new theory countenances makes no room for them.[86] Concomitantly, the quantities of total angular momentum and total energy, which in the original theory are primitive and therefore can take arbitrary values now are assigned a non-arbitrary value as a direct consequence of how the theory is conceived. This means that the new theory explains and offers insight into why these quantities take the value they do. Finally, the new theory allows us to diagnose why the pathological models arose in the original theory. In the case at hand, it is the unwarranted separation of kinetic and potential energy that is the cause. Thus, the example offers us a template for what an ideal instance of reformulation that eliminates unwanted models should look like. It also teaches us an important lesson: frequently, pathological models are not merely a fluke. The fact that they exist points to some significant shortcomings of the theory in which they occur.

3.2 Expansive Reformulation

For most of the discussion, we have focussed on reformulating theories in order to eliminate pathological models, in analogy with eliminating surplus structure. However, reformulation can also aim at including new models, in order to represent states of affairs that we deem to be physically possible but cannot adequately model in a theory as it is formulated. In these cases, we are not concerned as we were before with possibility inferences, i.e. the conclusion that a state of affairs is physically possible because it is consistent with the theory, but with impossibility inferences, i.e. the conclusion that a state of affairs is physically impossible because it is inconsistent with the theory. When we want to call such an impossibility inference into question because we suspect that a state of affairs is physically possible after all, we are motivated to reformulate a theory by the fact that presently, the theory's formalism does not allow us to represent that state of affairs.

A nice example of this can be found in the literature on "Newtonian Cosmology". In brief, the issue is this: prima facie, it seems as though it is physically possible for there to be

[86] This point is also well illustrated by the elimination of the standard of rest in the move from Newtonian to Galilean spacetime: in this case, a class of distinct models is collapsed into one and, importantly, this is because the new theory does not admit the kind of structure, viz. a standard of rest, that would allow for the distinction between the original theory's models in the first place. Thus, the original theory's distinction between models by a standard of rest is not even intelligible from the point of view of Galilean spacetime.

infinitely many massive bodies, distributed more or less homogeneously across space. Consequently, we would like to model this possibility in Newtonian Gravitation Theory. As it turns out however, this leads to technical difficulties if we are working within a mathematically crude formulation of NGT. We therefore find ourselves in precisely the kind of situation motivationalists are concerned with: our theory tells us something – in this case that there is no consistent mathematical description of an infinitude of gravitating bodies. We perceive this as a shortcoming of the theory, since it seems to us as though it should be possible for there to be a universe with a homogeneous non-zero mass distribution throughout. However, it is not known to us whether we can improve the theory so as to accommodate such a state of affairs. We are thus motivated to reformulate the theory, but we are prepared to accept that there could not have been infinitely many massive bodies according to NGT if this should turn out to be irreconcilable with the theory. That at least is the motivationalist attitude towards the scenario at hand. We will now look at the example in slightly more detail, but to spoil the result: as it turns out, Newtonian Gravitation Theory can be reformulated so as to describe uniform matter distributions in a consistent way. The best way to do so is to move to Newton Cartan Theory.

Here is how the story goes in more detail – I'm following the presentation in Malament (1995): In a simple formulation of Newtonian Gravitation Theory, we are describing things directly in terms of gravitational forces between objects, subject to the inverse square law. As Norton (1992) points out, this does not allow for a consistent description of a universe with a homogeneous, positive mass distribution. Such a mass distribution is best modelled via a constant mass density function ρ, representing the averaged mass density across the universe. Given such ρ, the force formalism yields for the gravitational force at a point r the following formula

$$\int dV' G\rho(r') \frac{\overline{rr'}}{|\overline{rr'}|^3} \tag{5.1}$$

For ρ constant and positive however, the integral does not converge. No consistent description of such a state of affairs is possible within the simple force formalism in which we are working.

To achieve a consistent mathematical description, we must move to a different formalism. In other words, we are motivated to reformulate the theory. One way to do so is to move to a formulation in terms of a gravitational potential ϕ. ϕ is governed by the Poisson equation

$$\nabla^2\phi = 4\pi G\rho \tag{5.2}$$

This equation now has solutions for constant positive ρ. They take the form

$$\phi(r) = \frac{2\pi}{3}G\rho|rr_0|^2 \tag{5.3}$$

where r_0 is an arbitrarily chosen point in space. So, we have achieved a reformulation of the theory that allows us to give a consistent description of a universe with an infinitude of homogeneously distributed gravitating bodies. From a motivationalist point of view, this now allows us to infer that such a universe is physically possible; something we were initially led to doubt when we were working in the simpler force formulation of NGT.

The reformulation we have found is not perfect. Given constant positive ρ, we find infinitely many non-isomorphic solutions, each of which arbitrarily privileges some spacetime point r_0. So, we cannot quite rest yet. But as Malament shows, these distinct solutions are empirically equivalent and arise because the gravitational potential is a gauge quantity. Through further reformulation, specifically through moving to Newton-Cartan Theory, we can eliminate this gauge freedom. Consequently, the reformulation maps each of the distinct solutions that we found for any given ρ to the same model of NCT, up to isomorphism. This latter part of the story is old news – it is an instance of reformulation with the aim of eliminating surplus structure. The first part of the story however – the move from the force formulation to the potential formulation – involves a source of motivation for reformulation that we have not seen before: the inability of a theory to consistently model a state of affairs that pre-theoretically we deem possible.

This concludes our sketch of motivationalism about nomic possibility. Much more remains to be said, of course. The most important question to be answered is whether a form of motivationalism about nomic possibility can be made plausible. Thinking back to Earman's discussion of the paradoxes of time travel, one might worry that any model of GR that leads to paradox can immediately be seen to be unphysical, even without reformulating the theory. This puts immediate pressure on motivationalism about nomic possibility as just presented. But this discussion will have to wait for another day. In the next Chapter, we consider a motivationalism put forth by Read and Moller-Nielsen [2020a], viz. motivationalism about dual theories.

6 Motivationalism and Duality

1 Introduction

The motivationalism we encountered so far was a view about the representational equivalence of models within one theory. But it naturally suggests an answer also to the question of when it is legitimate to regard two theories as equivalent, since there is a robust analogy between symmetries and theoretical equivalence: just as a symmetry is understood as a formal relationship between empirically equivalent models, a theoretical equivalence is a formal relationship between empirically equivalent theories, i.e. classes of models. Indeed, perhaps the most natural way to look at symmetries and dualities is the one suggested by Butterfield [2018]: symmetry is a special case of duality, viz. a self-duality to be found within a theory.

This analogy suggests, at least *prima facie*, that whatever preconditions motivationalists place on admitting symmetry-related models as representationally equivalent, they are also going to want to place on admitting formally and empirically equivalent theories as theoretically equivalent. In particular, motivationalists are going to demand an account of an ontology, or "metaphysical picture", that one can recognize as shared between two formally equivalent theories, before they are going to admit full equivalence between them. And indeed, a motivationalist perspective on theory equivalence has been defended by Read and Moller-Nielsen [2020a] for the case of dual theories.[87] In parallel with motivationalism about symmetries, they argue that it is not permissible *ab initio* to regard dual theories as equivalent. Only once a clear account of the ontology grounding their equivalence has been given is one justified to do so. The way to give such an account is to show that the dual theories share a common core.[88] I am in broad agreement with their view, although as before, I believe that insofar as finding a common core reformulation of the theories in question and giving an account of their shared ontology can come apart, it is the former that is decisive for judgments of equivalence. My aim in this chapter is to spell this form of motivationalism out in more detail and defend it against important objections.

[87] De Haro and Butterfield [2025] also consider themselves motivationalists about duality

[88] The importance of common cores to judgments of equivalence for dual theories is also emphasized by Read and Le Bihan [2018]

The structure of the present chapter is as follows: in Section 2, I introduce the notion of duality. Section 3 then puts motivationalism about dual theories on the table and surveys the various choice points a motivationalist of this kind faces. We will see that these by and large correspond to the points of contention that distinguished the versions of motivationalism about symmetry-related models we identified in Chapter 3. I will settle on a version of motivationalism according to which it is the existence of a viable common core theory that licenses us to regard duality-related theories as equivalent. In Section 4, we develop this version of motivationalism, and the notion of a common core in particular. Section 5 highlights and addresses a number of objections to the motivationalism defended here. Finally, Section 6 develops a distinction between vertical and horizontal reformulation in more detail. This distinction plays a role in the motivationalism this chapter defends.

2 Duality

The label 'duality' comes from string theory, where it means a systematic, uniform correspondence between the models of two theories that is understood to preserve empirical content. The most well-known examples are the AdS/CFT correspondence and T-duality.[89] What is striking about these theory pairs is that the dual theories offer seemingly interchangeable descriptions and yet involve at least apparently very different ontologies (see Read and Moller-Nielsen [2020], p. 278). This lends *prima facie* support to a motivationalist attitude towards them. In the case of T-duality for instance, dual models disagree over the radius of a certain compactified dimension of space. In the AdS/CFT case, the disagreement is even more stark, since the two theories ascribe a different dimensionality to space.

Abstracting from particular examples, let us make precise the formal features of dualities. I will adopt the definition used by Read and Moller-Nielsen [2020]. They characterize duality between two theories T_1 and T_2 as follows: first, let $\mathcal{D}_1$ and $\mathcal{D}_2$ be the spaces of DPMs of T_1 and T_2, respectively. Then T_1 and T_2 are dual just in case there is an isomorphism preserving empirical content between $\widetilde{\mathcal{D}}_1$ and $\widetilde{\mathcal{D}}_2$, where $\widetilde{\mathcal{D}}_i$ is a reduced version of $\mathcal{D}_i$

[89] For a philosophical introduction to dualities in string theory, see e.g. Rickles [2011]. For discussion of the AdS/CFT correspondence, see e.g. Horowitz and Polchinski [2006] or de Haro [2016]. T-duality is discussed inter alia in Becker et al [2007]

obtained via quotienting by any symmetries of T_i. To illustrate, NGT is dual to Maxwell Gravitation, since there is an empirical content-preserving isomorphism between $\widetilde{\mathcal{D}}_{MG} = \mathcal{D}_{MG}$ and $\widetilde{\mathcal{D}}_{NGT}$, where $\widetilde{\mathcal{D}}_{NGT}$ is obtained through quotienting by kinematic and dynamic shifts.

Let me emphasize that even though this definition of duality mandates quotienting by symmetries, it is not to be understood as presupposing that symmetry-related models are invariably physically equivalent. Rather, the purpose of requiring quotienting is merely to ensure that the definition of duality encompasses all of the equivalences of interest to the motivationalist. If we made duality dependent on there being an isomorphism between the unquotiented spaces of DPMs, MG and NGT would turn out not to be dual, for instance. Hence, the quotienting condition.

An alternative definition of duality has been proposed by de Haro and Butterfield (de Haro [2019a, b], de Haro and Butterfield [2021], [2025], Butterfield [2021]). But their understanding of duality is too restrictive for our purposes, since it makes it a definitional requirement that dual theories should share a common core. A definition of duality cannot prejudge the issue of whether all dual theories share a common core if it is to include the cases of interest to motivationalists however. After all, according to the motivationalist standpoint I want to defend, the equivalence of dual theories will depend on whether a common core can be found, which is supposed to be a non-trivial matter. So, it is preferable to work with Read and Moller-Nielsen's definition. That being said, an important aspect of Butterfield and de Haro's view is that dual theories can be inequivalent despite sharing a common core. This is *prima facie* in conflict with the view defended here. I will comment on the divergence below, when I consider objections to motivationalism.

3 Motivationalism about duality

We now want to develop motivationalism about dualities in detail. Our strategy will be the following: we will put on the table a very stringent form of motivationalism, according to which dual theories may only be interpreted as equivalent if certain strict conditions are met. Next, we will outline a number of objections to motivationalism and see how the very strict doctrine we use as a starting point must be transformed and weakened to

accommodate these objections. The main contention of this chapter is that even after certain concessions are made, there remains a plausible robust and clearly defined form of motivationalism. Our discussion will be accelerated by the fact that many of the considerations that shaped the motivationalism about symmetry-related models we defended in Chapter 3 will bear directly or indirectly also on motivationalism about dual theories. The strict motivationalism about dualities that we will treat as our starting point comprises the following three tenets:

First, "one needs to give a coherent account of the shared ontology (i.e. an appropriate interpretation) of duality-related models, before one declares those models to be physically equivalent; moreover, there appears to exist no set of a priori principles by which one may deductively infer that such an interpretation exists." (Read and Moller-Nielsen [2020a, p. 284])

Secondly, the way one must give this coherent account of the shared ontology is to reformulate the dual theories T_1 and T_2 so as to lay bare that they share a common core (Read and Moller-Nielsen [2020a]). More precisely, the idea is this: since T_1 and T_2 are dual, each model M of T_1 corresponds to a dual model M' of T_2. If T_1 and T_2 are suggestive of incompatible ontologies, we cannot simply interpret them as equivalent: rather, it must be shown that the appearance of incompatibility is entirely due to differences in the representational artifacts employed by M and M', respectively. Each of M and M' must therefore be reformulated (in a way to be made precise) and the resulting models – call them M* and M'*, respectively, must be isomorphic. M* is called the common core of the dual models M and M'. Moreover, the set of common cores M* for all models M of T_1 forms a theory in its own right, which we can call T^*. T^* is the common core of the dual theories T_1 and T_2. According to strict motivationalism then, we are only permitted to interpret T_1 and T_2 as equivalent once a common core T* has been constructed. That being said, of course the way T^* is to be constructed is yet to be spelled out. The paradigm here is the construction of Newton-Cartan Theory from NGT (see Chapter 4), but we will say slightly more about core construction in the abstract in the next section.

Thirdly, as a corollary of the second tenet, if M and M' are already isomorphic, no reformulation is required. However, in general, this will not be the case: "dual models are generically not isomorphic: straightforwardly understood, they represent worlds which

differ more than purely with regard to which particular objects are playing which qualitative roles" (Read and Moller-Nielsen [2020a], p. 278).

Finally, we are going to want to insist on an analogue of Martens' [2018] 'no defeaters' condition (see Chapter 3). This condition will become that a core can only warrant a judgment of equivalence for the models or theories to which it is common if it is not explanatorily deficient relative to these superstructures. The canonical example here is electrodynamics in the vector potential formulation, with its core of electrodynamics in the Faraday tensor formulation. As we noted, the vector potential formulation has more explanatory power than the Faraday tensor formulation, which prevents us from interpreting SRMs of vector potential ED as equivalent on the basis of the Faraday tensor core. This 'no defeaters' condition carries over to the common core account of equivalence for dual theories. That being said, I do want to make room for there being an interpretation of vector potential ED according to which it is a vehicle for expressing Faraday tensor ED, or, in the language of de Haro and Butterfield [2021], a representation of Faraday tensor ED. In other words, while the mere fact that Faraday tensor ED is a core of vector potential ED does not warrant viewing the two as equivalent (in light of the differences in explanatory strength), it does give licence to those who embrace Faraday tensor ED to treat vector potential ED as a formulation of the theory they accept.

4 Common Cores

The most worked out account of common cores is due to de Haro and Butterfield (de Haro and Butterfield [2021], Butterfield [2021]). In their version, the notions of a bare theory and of what they call specific structure plays a crucial role: the bare theory is supposed to be the artifact-free core of the theory, whereas the specific structure is supposed to be the representational fluff, i.e. gauge structure. Butterfield illustrates the idea with the example of Newtonian spacetime: Galilean spacetime is the bare theory, the standard of rest is specific structure. While we will be broadly guided by de Haro and Butterfield's conception of common cores, I don't find this particular distinction all that helpful for a number of related reasons. First, nothing internal to a theory distinguishes specific structure from core structure – there is no big red sign saying: this is conventional or surplus. Relatedly, different interpretations of a theory will correspond to different ways of drawing the

core/specific distinction.[90] Take Butterfield's [2018] example of Galilean spacetime: he treats it as a bare theory relative to Newtonian spacetime, but of course one could also treat Galilean spacetime as a theory with specific structure whose core is the bare theory of Maxwell spacetime. In what follows, I will therefore not rely on these notions.

Recall that according to motivationalists, we are permitted to interpret symmetry-related models of a theory T as representing the same state of affairs only if there exists a different theory T* such that there is a surjective mapping $f: T \to T^*$ with the property that whenever models m and n of T are symmetry-related, f(m) is isomorphic to f(n). Moreover, f(m) and f(n) must capture the "common core" of m and n, i.e., they must be definable from m and n, respectively, and free of the surplus structure that distinguishes m from n. The paradigmatic example is $T = NGT$ and $T^* = NCT$, with f defined via the Trautman recovery construction. Thus, if m and n are symmetry-related models of NGT, there exist isomorphic models m* and n* of NCT that capture the "common core" of m and n. We will call T* a *sophistication* of T and f the *recovery function*.

We can now extend this account to dual theories. Let T_1 and T_2 be dual, with $f: \widetilde{\mathcal{D}}_1 \to \widetilde{\mathcal{D}}_2$ the empirical content-preserving isomorphism between $\widetilde{\mathcal{D}}_1$ and $\widetilde{\mathcal{D}}_2$. Then T_1 and T_2 can be interpreted as equivalent just in case there exists a theory T^* (the common core theory), which is a sophistication of both T_1 and T_2, with recovery functions $h: T_1 \to T^*$ and $j: T_2 \to T^*$ such that the following condition is met:

- For all $[m] \in \widetilde{\mathcal{D}}_1$: Whenever $m' \in [m]$ and $n \in f([m])$, $h(m') \cong j(n)$

Note that this entails the following:

- For all $[m] \in \widetilde{\mathcal{D}}_1$: for all $m_1, m_2 \in [m]$, $h(m_1) \cong h(m_2)$
- For all $[n] \in \widetilde{\mathcal{D}}_2$: for all $n_1, n_2 \in [n]$, $j(n_1) \cong j(n_2)$

The most impressive and fruitful applications of the common core account can be found in discussions of what is called the geometric trinity, in both its relativistic and non-relativistic form (March, Wolf and Read [2024]). The relativistic geometric trinity is given by the theories of general relativity (GR), teleparallel gravity (TG) and symmetric teleparallel gravity (STG) (Capozziello, de Falco and Ferrara [2022]). All three theories share the same basis for their KPMs, which takes the form $\langle M, g_{ab}, \Phi \rangle$, with M the base manifold, g_{ab} the Lorentzian metric, and Φ a representation of the matter field. Where the theories come

[90] De Haro and Butterfield are aware of this, of course (see de Haro and Butterfield [2021], p. 2983)

apart is the choice of a connection (March, Wolf and Read [2024]). For GR, the connection is the Levi-Civita connection, which is compatible with the metric and torsion-free, but in general curved. The connections of TG and STG are each constructed from the Levi-Civita connection via addition of a tensor. In the case of TG, that added tensor is chosen in such a way that the resulting connection is flat and compatible with the metric but torsionful. For STG, the resulting connection is not compatible with the metric but flat and torsion-free. The three theories share the same dynamics, as captured by the Einstein field equations. They are therefore observationally equivalent. A judgment that they can be interpreted as fully equivalent is supported by the fact that GR constitutes their common core in a precise sense (March, Wolf and Read [2024]).

As it turns out, there is a non-relativistic analogue to this geometric trinity. The three theories it comprises are the non-relativistic limits of GR, TG, and STG, respectively (Wolf, Read and Vigneron [2023]). In the case of GR, the non-relativistic limit is Newton-Cartan Theory. These three theories share as their common core Maxwell Gravitation, which has less structure than each of them. (March, Wolf and Read [2024]).

Having seen the common core account in action, several clarifications are in order, regarding what the account says and what it doesn't say: First, take the common core relationship afforded by electrodynamics: Let T_1 be the vector potential formulation of electrodynamics, and $T_2 = T^*$ the Faraday tensor formulation.[91] This is of course a degenerate case since $T_2 = T^*$. Nonetheless, it is an instance of theories that share a common core. To reiterate a point made in the previous section, our account of equivalence does not say that electrodynamics in the vector potential formulation is equivalent to electrodynamics in the Faraday tensor formulation. It does say that there is a permissible interpretation of vector potential ED according to which it is equivalent to Faraday tensor ED.[92] Secondly, as of now at least the account does not say that the common core theory T*

[91] An interesting and surprising non-example is electrodynamics in the fibre bundle formulation as compared to electrodynamics in the Faraday tensor formulation. Since these are commonly thought of as, respectively, the canonical sophistication and reduction of electrodynamics in the vector potential formulation, one would expect that Faraday tensor ED is a core of fibre bundle ED. But this turns out not to be the case! In fact, as Weatherall [2018] points out, there is a robust sense in which fibre bundle ED posits less structure than Faraday tensor ED. One way in which this manifests is that the functor from the category of Faraday tensor ED to the category of fibre bundle ED is not full, and so not a categorical equivalence. The wider implication of this example is that contrary to what one might expect, the canonical reduction of a theory is not invariably a core of the canonical sophistication.

[92] This verdict is congenial to Coffey's [2014] account of equivalence as interpretative equivalence.

must be “artifact-free”, “ontologically pristine”, “fundamental”, or anything of the sort. Adding this condition would render the account more similar to Sider’s [2020] account of equivalence. Whether it is reasonable to impose such a demand will be discussed in more detail below.

In the next section, we will subject motivationalism about dualities to scrutiny. But before we turn to criticism, let us first appreciate some of the strengths of the common core account of equivalence. One of its major advantages is that it accommodates the insights that motivated the categorical equivalence programme while avoiding the latter’s flaws. One driver of the categorical equivalence account was the observation that theories can fail to be interdefinable and yet be viewed as equivalent, both in the practice of physicists and also in the eyes of the more careful, reflective philosopher. This can happen when one of the two theories introduces gauge structure not present in the other theory (Weatherall [2016b]). A good example of this is the theory of electromagnetism in the vector potential formulation as opposed to the Faraday tensor formulation: the former, unlike the latter, exhibits gauge structure. In consequence, a multitude of non-isomorphic models of vector potential electrodynamics will correspond to a unique model of electrodynamics. Nonetheless, it is standard to treat the two theories as equivalent, or to at least maintain that the vector potential formulation has an interpretation according to which it is equivalent to the Faraday tensor formulation.

It is this state of affairs that the category theoretic account of equivalence was meant to capture (see [Weatherall 2016b]). However, one notes that it does not do the job it was supposed to do very well. If we equip each of vector potential electrodynamics and Faraday tensor electrodynamics with the most natural choice of category, in which arrows between models are fully structure-preserving maps, the two will turn out not to be equivalent! To remedy this, proponents of categorical equivalence usually suggest that we associate a different category with vector potential electrodynamics (see e.g. Dewar and Eva [unpublished]). In that alternative category, the arrows are less than fully structure-preserving, as a result of which any two SRMs of vector potential electrodynamics are rendered isomorphic. This move has the air of being ad hoc however. The category theoretic account of equivalence struggles also with a host of other problems, the preeminent of which is that the category associated with a theory usually fails to capture all of the theory’s models’ internal structure, which manifests in categorical equivalences

between theories that are not intuitively equivalent.[93] In other words, it is often thought that categorical equivalence is too weak a standard of theoretical equivalence (see e.g. Barrett and Halvorson [2016b]).

The common core account does better: it allows us to interpret vector potential electrodynamics as a form of Faraday tensor electrodynamics on the entirely natural grounds that the latter is a core of the former. Importantly, it is not necessary to make an unfounded assumption that any two models of vector potential electrodynamics related by a gauge transformation are equivalent to reach this verdict, and in consequence, there is no need to pretend that gauge-equivalent models are isomorphic, as advocates of the categorical equivalence standard are wont to do. By contrast, if we want to reach the verdict of equivalence while operating with this categorical equivalence standard, we have to bake such an assumption into our choice of category for vector potential electrodynamics, without any further explanation as to why it should be warranted. The most natural choice of category for vector potential electrodynamics does not deliver the right verdict, as we noted.

5 Motivationalism about duality examined

We have presented motivationalism about dualities. Now let us examine it. But before we consider genuine objections to motivationalism, let me briefly comment on a merely apparent one. One view that seems to be in conflict with the motivationalism defended here is Butterfield's [2018]. He maintains that dual theories can be inequivalent despite sharing a common core. One example he gives is that of two models of NGT (on Newtonian spacetime) related by a kinematic shift. The models share a common core, which according to Butterfield is given by the unique model of NGT on Galilean spacetime embedded in both of them. Nonetheless, Butterfield maintains that the two models can be interpreted as disagreeing. He emphasizes that Newton would have interpreted them as inequivalent, and that Newton's position was a respectable one, especially since no viable alternative to NGT

[93] See primarily Weatherall [2021] for an authoritative survey of problems . But see also Dewar [2023] for some pushback against the idea that the category associated with a theory erases too much of the internal structure of the theory's models, and March [forthcoming] for a contribution towards reviving the categorical equivalence programme.

was on the horizon at the time, so that it was far from clear how the two models could possibly be equivalent.

I agree! Butterfield's view is not in conflict with the motivationalism defended here. It will be instructive to see why. According to the motivationalist, the fact that the two models share a common core means that one is *permitted* to interpret them as equivalent. However, an alternative, "literal" interpretation, according to which the two models are inequivalent remains admissible, too. The two standpoints are two inequivalent ways to interpret the formalism of NGT: according to one, the standard of rest of NGT is physical; according to the other, it is gauge (see Coffey [2014]). However, while both interpretations are possible, it remains the case that if they had been available at Newton's time, it would have been rational to believe that the non-literal one tracks the truth, and not the literal one. With this clarification out of the way, let us now consider genuine objections to motivationalism about dualities.

5.1 A host of problems

As we will see, motivationalism faces a tangle of issues. These mainly concern the question of whether judgments of equivalence for dual models invariably require a common core and a coherent account of the ontology shared by the models. We will look at a number of cases of theory comparison that seem to cause problems for motivationalism, with the aim of isolating a few different objections. But while we are going to distinguish several concerns for the sake of clarity, it will be worth bearing in mind that they are deeply interlinked.

First difficulties for motivationalism arise when we consider cases in which judgments of equivalence between theories seem to precede a settled interpretation of the theories in question. The most important case of this is that of quantum mechanics. As is well-known, this theory exists both in the Schrödinger and in the Heisenberg formulation. Von Neumann [1932] showed to virtually everyone's satisfaction that these are fully

equivalent.[94] The correct interpretation of the theory however remains a matter of much controversy. This leads to two problems for motivationalism.

The first is this: we seem to be dealing with a case in which we can know that two theories are equivalent even without possessing an account of the causal-realistic metaphysical story grounding their equivalence. That is so simply because we don't have any clear metaphysical interpretation of the quantum formalism; or at least we didn't have one at the time von Neumann published his proof. Much of the interpretive work on quantum mechanics, such as the development of the Everett interpretation, came much later. So, it seems as though motivationalists must either admit that in some cases, we can know that two dual theories are equivalent despite our lack of a coherent account of their shared ontology, or dig their heels in and insist that von Neumann's proof did not suffice to show that the Schrödinger and Heisenberg formulations are equivalent. Neither option seems terribly appealing. Call this the *'equivalence without interpretation'* objection to motivationalism. It is a problem for anyone who sees judgments of equivalence as grounded in interpretation rather than agreement of mathematical structure and was identified in this form *inter alia* by Coffey [2014].

A second, and even more pressing problem for motivationalism arises because the example is one in which the equivalence is established not by constructing a common core but by finding a mathematical correspondence between the two classes of models. This puts pressure also on the motivationalist claim that dual theories may only be interpreted as equivalent once we have found a common core. Call this the *'equivalence without common core'* objection. This second objection can also be supported by other examples. As David Wallace [2022a] has pointed out, there are many different theories of Euclidean space for instance. One can define Euclidean space via a class of preferred coordinatizations, or as an affine space, or as a manifold with a flat non-degenerate metric, etc. These definitions are mathematically distinct. Nonetheless, there is widespread consensus that they are equivalent, and crucially, this consensus is not rooted in the existence of a common core. Wallace addresses this explicitly:

> It would miss the point to argue that in this case we need some intrinsic characterization of Euclidean space underlying all of these accounts (a Tarskian axiomatization, say). The norms of math and physics regard the coexistence of different accounts of Euclidean space as innocuous; hence, [...] the 'we' who need

[94] At least they are equivalent after some rational reconstruction. As Muller [1997] has shown, the original versions of these theories were inequivalent.

> that intrinsic characterization are not those trying to construct theories of physics. The need for that axiomatization would come, if at all, from some further philosophical project [...]
>
> (Wallace [2022a], p. 354)

The challenge for motivationalists is to defend the idea that common cores are crucial for judgments of equivalence against these apparent counterexamples. Let me mention one further example in this vein, which will then also lead us to a third objection to motivationalism. This example has been discussed at length by Barrett and Halvorson [2016]: as is well-known, various geometries can be axiomatized in a way that treats lines as basic or alternatively in terms of points as the primitive entities. Fix some geometry, and call the corresponding axiomatizations in terms of lines and points T_L and T_P, respectively. Once again, we face a pair of theories that are widely regarded as equivalent even though there is no obviously satisfactory common core. They possess a common extension, in which both points and lines are explicitly represented. One might worry however that this common extension does not meet the motivationalist's requirements for a common core since it is not ontologically perspicuous, as witnessed by the fact that in representing both points and lines as primitive, it invariably involves representational redundancy – or so the objection goes. By the criteria of simplicity and parsimony, the common extension fares worse than either of T_L and T_P. So, if anything, T_L and T_P are our best candidates for the common core, if we insist that the common core should give us a coherent account of the ontology shared by the two theories. But it seems impossible to decide which it is. Indeed, many would deny that any matter of fact adjudicates which choice is right.

This leads to a third objection to motivationalism. Its target is the assumption implicit in the motivationalist doctrine that for any theory, we can specify a unique ontology that unequivocally comprises "what the theory says there is", as it were – that is then the ontology the common core is supposed to present. The worry about this assumption raised by the case of T_L and T_P is that the two theories appear to involve different ontologies and that it might be in principle undecidable which of the two is to be preferred. Much ink can be spilled over the question of which of the two is more ontologically correct, or perspicuous, or fundamental, or natural, etc. But since it is hard to see how to make progress on this question, the impression arises that the debate is fruitless and misguided, and rests on the false presupposition that there must be some privileged version of the theory that gets at the "real" ontology. The alternative is that equivalent theories can involve different, and apparently incompatible, specifications of an ontology, but that there is no

sense in which one of these is privileged over the other. Putnam [1977, 1980, 1987] calls this view 'conceptual relativity'. Motivationalists, it seems, are implicitly committed to a denial of conceptual relativity. After all, one of their central claims is that judgments of equivalence for theories must proceed via a common core theory that presents a coherent account of the ontology common to the two theories. Cases such as that of T_L and T_P put pressure on the presupposition that such a coherent account is always to be had.

Call this the *'scholastic metaphysics'* objection to motivationalism. The label is due to the charge that motivationalists must think of theory interpretation as the act of specifying a unique ontology of objects countenanced by the theory and properties had by those objects. The 'scholastic metaphysics' objection is another objection that we see echoed in Wallace. He states that "a predicate precisification of a mathematically-given physical theory is a presentation of that theory [...] in terms of objects, their properties and their relations. Then given predicate precisifications of mathematically-equivalent theories, there is often no simple relation of the objects in the first to the objects of the second; likewise the properties and relations. Mathematically-equivalent theories cannot just be construed as talking about the same entities or ascribing the same properties to them." (Wallace [2022a], p. 356) One example he gives is that of different but equivalent presentations of a scalar field theory, which differ over whether they treat the field as a separate entity from the spacetime on which it lives. But as he emphasizes, the phenomenon is pervasive in physics.

To summarize, we have identified three, albeit closely interrelated objections: the 'equivalence without interpretation' objection, the 'equivalence without common core' objection, and the 'scholastic metaphysics' objection. In what follows, I will aim to develop a version of motivationalism that can accommodate these objections. While I will address the third to some extent, I will here mainly deal with the first two. The subsequent chapter looks at the 'scholastic metaphysics' objection in more detail. I will also dedicate a subsection of the discussion to recent work by Grimmer et al. [forthcoming]. They have argued that even in simple cases, dual theories do not always possess a common core. This exacerbates the 'equivalence without common core' problem.

5.2 A defence of motivationalism

How can motivationalists respond to the objections just assembled? Some of the objections suggest that we can know that two theories are equivalent even though we are not agreed on the theories' interpretation, and that motivationalism is wedded to an antiquated scholastic metaphysics according to which interpretation involves finding a unique correct ontology for a theory. The way to accommodate these points, I will argue, is to weaken the common core account: roughly, the view motivationalists should take is that while judgments of equivalence must be based on the construction of a common core, that common core need not already possess a clear interpretation, nor need it be without "rival" common cores that seem to display a different ontology. I will elaborate on this shortly.

However, there is also the worry that we can in certain cases know that two theories are equivalent prior to having found a common core. This suggests that the common core account should be abandoned entirely, contrary to what I have just proposed. In response, I will aim to show that motivationalists can stand firm on insisting on the common core account of equivalence. I will proceed as follows: first, I will explain how the common core account can be weakened so as to accommodate some of the insights to be gained from the objections to motivationalism, and I will argue that much speaks in favour of retaining it in this weakened form. Then, I will explain how we can do so despite the 'equivalence without common core' objection.

Its philosophical underpinnings make the common core account of equivalence an eminently natural choice for motivationalists. This proximity is made especially evident in McSweeney's [2016] work on theoretical equivalence, which can be read as an early development of something like a common core account, albeit not under this label. We can turn to McSweeney for help with finding a more moderate motivationalism that avoids the 'scholastic metaphysics' and 'equivalence without interpretation' objections, but crucially without giving up the insistence on having to find a common core theory. I will first give an overview of her work, so that I can subsequently explain how this can be done.

First, it is noteworthy that McSweeney emphasizes, as I have been aiming to do, that many of the important questions about theoretical equivalence concern the epistemology of equivalence. Thus, her aim is not to tell us what equivalence is but to identify conditions

under which one is warranted in judging two theories to be equivalent. But the affinity between motivationalism and her early common core account is not merely a matter of emphasis but also a matter of substance. Thus, she writes that "in order to be justified in believing that two theories, T and T', are equivalent, there must be an occupiable perspective from which T and T' can be conceived of as a single unified theory, T+, which (in some to-be-determined sense) says nothing over and above either T or T', and which says everything that T does and T' does." (McSweeney [2016], p. 270)

One important part of McSweeney's condition that requires some unpacking is the phrase "occupiable perspective". Of course, in the abstract motivationalists will take this perspective to be offered by the common core theory, but to see more concretely what is meant by this, it will be helpful to recall the example of Gravitation Theory on Newtonian spacetime discussed in Chapter 3. This example gives us quite a good intuitive handle on what an "occupiable perspective" could look like. In that theory, recall, we find the symmetry of Galilean boosts. Moreover, we noted that we are justified in regarding models related by this symmetry as equivalent since we can set our Gravitation Theory on Galilean spacetime, thereby eliminating the structure that distinguishes between models related by Galilean boosts. So, NGT on Galilean spacetime offers an "occupiable perspective" from which models of NGT related by a boost can be interpreted as equivalent.

Let us now highlight two features of the move from Newtonian spacetime to Galilean spacetime that characterize more generally what constitutes an "occupiable perspective" for the more moderate motivationalism I want to propose. First, the theory on Galilean spacetime is no worse than the theory on Newtonian spacetime: the move has not resulted in an explanatory loss. This is not what we find in other examples, such as the move from vector potential electrodynamics to loop holonomy electrodynamics for instance. Here, the new theory cannot explain everything the old theory can and therefore must resort to positing cosmic conspiracies, as we saw. Nothing like that occurs in the move from Newtonian to Galilean spacetime. This marks one important characteristic of the kind of "occupiable perspective" that allows us interpret dual theories or symmetry-related models as equivalent: the perspective must be given by a theory that is no worse than either of the dual theories or SRMs we intend to regard as equivalent. After all, if the unifying perspective can only be obtained at a loss of, say, explanatory power or elegance, it will be preferable to uphold the judgment of inequivalence.

The second aspect of the example worth highlighting is that the "occupiable perspective" offered by Galilean spacetime is not one that can claim to be perfectly natural or fundamental. As Knox [2014] has argued, Gravitation Theory set on Galilean spacetime exhibits symmetries of its own, so that Galilean spacetime cannot be the appropriate setting for Newtonian Gravitation Theory. Rather, one must move to Newton-Cartan Theory. Galilean spacetime is merely a "halfway house". But what is worth noting here is that despite this, the theory still warrants our interpretation of boost-related models as equivalent. An occupiable perspective that supports judgments of equivalence need not be a perfect perspective. Thus, the phrase "occupiable perspective" incorporates a salutary moderation: if we are to judge two theories to be equivalent, we need a new theory that is no worse and from the point of view of which the two original theories are oblique descriptions of the same reality. But crucially, this new theory need not be the final word. It can retain some representational redundancy for instance, or it can fail to be perfectly perspicuous, and still support the judgment that the two original theories are equivalent.

Importantly, constructing a common core does not in general suffice for providing "a coherent picture of the common ontology of the pairs of models of these two theories related by the duality" (Read and Moller-Nielsen [2020], p. 280). The common core may be every bit as mysterious and unamenable to interpretation as the dual theories. In particular, the common core may itself contain residual surplus structure, and it may be unclear which of its structure is directly representational of physical structure – recall the case of quantum mechanics: there is a debate over whether the wavefunction is representational or epistemic. So, while motivationalists frequently equate finding a common core with providing a coherent picture of the shared ontology, the two can come apart. What is crucial for judgments of equivalence is the former, not the latter: a common core can underwrite a judgment of equivalence of dual theories even absent a metaphysical interpretation of any of the theories involved.

In brief then, my suggestion is to hold on to the common core criterion, but to give up the condition that the common core must be intrinsic, or surplus-structure free, or ontologically perspicuous, or without interpretive challenges. This of course brings us to the next challenge to motivationalism, i.e. the 'equivalence without common core' objection. Recall the example: von Neumann's [1932] equivalence proof for wave and

matrix quantum mechanics seems to establish equivalence without the provision of a common core.

One move motivationalists might be tempted to make in response is to suggest that von Neumann's result establishes an isomorphism between each matrix model and the corresponding wave model. This would allow them to explain why a common core is not needed here: we already noted when presenting motivationalism that a common core is only required if the dual models are not isomorphic. While I think that this is essentially the correct way for motivationalists to go, it is worth emphasizing that the correspondence between a matrix model and a wavefunction model deviates starkly from the paradigms of isomorphic models motivationalists have in mind when they suggest that isomorphic models can straightforwardly be interpreted as equivalent. Here, they mainly think of Leibniz shifts and the kinds of models we find in the hole argument – models, that is to say, that intuitively result from simply changing which manifold points are selected to represent which spacetime points (Moller-Nielsen [2017], Read and Moller-Nielsen [2020a]). In those cases, the ontology of the two models is recognizably the same, so that the standard solution of interpreting the models as equivalent on the basis of anti-haecceitism is reasonably plausible. Not so in the case of the correspondence between a matrix model and a wave model: here, it is not intuitively possible to interpret the models as differing merely with regard to which objects play which qualitative roles. At least *prima facie*, the models involve starkly different ontologies.[95] As Coffey states, "only one formulation (the Schrödinger one) seems to represent the state of a quantum system as dynamically evolving in time. Insofar as we expect an interpretation of a foundational formalism to tell us about which fundamental physical features dynamically evolve in time, and I suspect we do, this suggests that taking both quantum mechanical formalisms as essentially the same for the purposes of interpretation is mistaken." (Coffey [2014], p. 837n35)

The way out for motivationalists I would suggest is to treat certain mathematical equivalences such as the one established by von Neumann as sufficient for equivalence despite the fact that they relate theories whose ontologies superficially seem to differ. My suggestion here goes hand in hand with what I proposed above, viz. that since finding a common core does not invariably provide a perspicuous ontological picture, it is the existence of a common core, rather than that of a "coherent ontology" that is crucial for

[95] This point is emphasized by North [2020], Ch. 7

judgments of equivalence. In each case, the importance of telling an ontological story that links the putatively equivalent theories is relegated.

This allows us a new perspective on when a common core is required in order for a judgment of equivalence to be warranted. The motivationalist credo has it that reformulation into a common core theory is necessary whenever the dual or symmetry-related models are not isomorphic. So, only when two models are isomorphic are we permitted to interpret them as equivalent without having to reformulate them or find their common core, both in the intra-theoretic case – the case of symmetries – and in the inter-theoretic case – the case of duals. But here, motivationalists ought to weaken their stance. While isomorphism may be the appropriate standard in the intra-theoretic case, in the inter-theoretic case it is too strict. I take this to be the lesson from the example of point and line geometry and from Wallace's [2022b] aforementioned example of different definitions of Euclidean space (as an affine space, a manifold with a flat metric, etc.). In Halvorson's words, "two classes of models and M and M' might be equivalent even when there is no sense in which individual models in M are isomorphic to individual models in M'" (Halvorson [2012], p. 196).

(Classes of) non-isomorphic models can stand in a variety of relations that make it legitimate to regard them as equivalent. Among these are for instance bi-interpretability and Morita equivalence, but also other relations.[96] The correct criterion for when we are permitted to interpret two models as equivalent even absent a reformulation or common core should therefore be "equivalence by the standards of mathematics: a 1:1 transformation between models that preserves mathematical structure." (Wallace [2022a], p. 353) As Wallace concedes, this standard of equivalence, which I will call 'mathematical equivalence' going forth, is not fully precise. Indeed, the lack of a clear definition of mathematical equivalence has been flagged as a major obstacle to spelling out theoretical equivalence in terms of mathematical equivalence. Glymour [1977] proposes that two models are to be regarded as mathematically equivalent just in case the geometrical objects of one possess a unique covariant definition in the other, and vice versa. But as Weatherall [2016b] notes, Glymour's criterion fails to pin down a unique notion of interdefinability for the kinds of models we find in ordinary theories of physics. So, there will be room for disagreement in many cases, and more work needs to be done to give a

[96] I will defend the legitimacy of regarding Morita equivalence as such a standard of mathematical equivalence at length in the next chapter.

complete account of mathematical equivalence. To do so goes beyond the scope of this thesis however. These issues notwithstanding, motivationalists should embrace mathematical equivalence as their standard of equivalence. In other words, motivationalists should maintain that reformulation is necessary in order to interpret two models or theories as equivalent just in case the models or theories are not mathematically equivalent.

What is crucial now is that motivationalists can accept mathematical equivalence instead of strict isomorphism as their standard of sameness without giving the game away, i.e., without rendering their position vacuous. After all, there will be many models and theories that we might like to interpret as equivalent despite the fact that they fail to be mathematically equivalent by any reasonable standard. Take for instance vector potential electrodynamics and Faraday tensor electrodynamics. These fail to be mathematically equivalent simply because they are not related by the kind of 1:1 transformation Wallace speaks of – the relation between their models is many-one. Nonetheless, it should be admissible to have an interpretation of vector potential electrodynamics as Faraday tensor electrodynamics in disguise (although that is not the only, and arguably not even the best interpretation).

My response to the 'equivalence without common core' objection is therefore this: in the cases the objection invokes, we do indeed legitimately judge that the theories are equivalent absent a common core. But the motivationalism I defend can accommodate this: we are dealing with a case in which the two theories are mathematically equivalent. In such cases, a common core is not required for a judgment of equivalence. But crucially, to concede this is not to concede that we are permitted to interpret mathematically *in*equivalent dual theories as equivalent without first having found a common core. A principled distinction between these two kinds of cases can be drawn, which means that a well-delineated motivationalist view remains even once we widen our standard of evident equivalence from strict isomorphism to mathematical equivalence. Some might be skeptical about this claim, and so I will defend it at greater length in the final section of this chapter. First however, I will round off the present discussion by defusing some worries raised by Grimmer et al. [forthcoming] about the common core account of equivalence. That will conclude my discussion of the 'equivalence without interpretation' and

'equivalence without common core' objections to motivationalism. This leaves us with the 'scholastic metaphysics' objection. I will address it in the next chapter of this thesis.

5.3 A good heart these days is hard to find?

Recently, it has been argued by Grimmer et al. [forthcoming] that even in certain comparatively simple cases of duality a common core cannot be found, at least if we think of common cores in the strict sense of a theory constructed exclusively in terms of structure definable in both of the dual theories. This raises the worry that the common core account of equivalence does not deliver the required judgment in these cases.

Grimmer et al. focus on certain simple quantum mechanical scalar field theories. The duality in their examples is given by the Fourier transform. Here is their central example: Let T_{old} be a theory of a complex valued scalar field on a two-dimensional manifold $\mathcal{M} \cong \mathbb{R}^2$ representing one space and one time dimension. The Schrödinger equation for the system described by T_{old} is

$$i\partial_t \phi_{old}(t,x) = (-\partial_x^2 + x^2)\phi_{old}(t,x) + \frac{\lambda}{2}\big(\phi_{old}(t,x+a) + \phi_{old}(t,x-a)\big) \tag{6.1}$$

which corresponds to a Hamiltonian

$$H_{old} = -\partial_x^2 + x^2 + \lambda \cosh(a\partial_x) \tag{6.2}$$

They claim that a dual theory T_{new} can be obtained from the Fourier transform of H_{old}. The transformed Hamiltonian is

$$H_{new} = q^2 - \partial_q^2 + \lambda \cos(aq) \tag{6.3}$$

They take this as the Hamiltonian for the theory T_{new}, which is also a theory of a scalar field on a two-dimensional manifold. Crucially, however, that manifold is also supposed to be interpreted as a representation of *spacetime*, not of momentum space-time. This is what makes T_{new} a theory distinct from T_{old}.

Grimmer et al. make several claims about this theory pair. For one, they argue that a common core cannot be specified in terms of structure shared by the two theories. In order to obtain a theory T^+ that unifies the dual pair one must involve structure that goes beyond what can be found in either of the theories – what they have in mind here is a Hilbert space. This leads to a further worry: the unifying theory T^+ is supposedly based on an ontology of

state vectors. But such an ontology may be less desirable than the ontologies of T_{old} and T_{new}. If however these two theories are each superior to T^+, the existence of T^+ does not warrant an interpretation of T_{old} and T_{new} as equivalent. After all, such an interpretation would construe the two as "overwrought" versions of T^+ in order to account for their seeming inequivalence. If a literal interpretation of T_{old} and T_{new} is theoretically preferable however, we are deprived of the grounds on which we would judge them equivalent.

For defenders of a common core account of equivalence, there are three separate concerns here: first, that something like a common core will in many cases only be available if we admit as common cores theories the construction of which involves not only structure shared by the two theories but instead also structure that goes beyond what can be defined in both theories, and perhaps even in either theory. Of course, in such cases, the label 'common core' is somewhat misleading. 'Common extension' or 'common completion' would be more appropriate. Second, that in consequence the existence of a common core even in this extended sense is a non-trivial matter and not guaranteed even in the case of intuitively equivalent theories. And finally, third, that since common cores in this extended sense in general involve an ontology radically different from that of the dual theories, they do not provide grounds for interpreting dual theories as equivalent even when they exist.

If the example given by Grimmer et al. really is one in which the existence of a genuine common core is tenuous and where the candidate we can find is not apt to underwrite a judgment of equivalence between the two theories, I take it that this spells trouble for the common core account of equivalence. Of course, its defenders could simply dig in their heels and maintain that the two theories in question are dual but not equivalent. But that is a tough bullet to bite. Alternatively, one may want to regard the common core criterion as merely sufficient but not necessary for theoretical equivalence.

Luckily, I don't think defenders of a common core account of equivalence need to resort to either response. I want to make two points about the argument Grimmer et al. present. The first is that the example of a putatively problematic duality they give poses no challenge to the common core account of equivalence. Depending on how one fleshes out the example, the two theories in question are either dual but equivalent in the completely banal way in which theories that differ merely by a relabelling of entities are equivalent – recall Sklar's example of calling lions 'tigers' and tigers 'lions' – or they are simply not dual. The second

point I want to make (Section 2) is that when we consider the genuine duality between quantum theories in position and momentum representation, we *can* find a common core that licenses a judgment of equivalence. This judgment aligns with Butterfield's analysis of the case.

Let us first look at the pair of theories Grimmer et al. allege to be dual, i.e. T_{old} and T_{new}. In the story they tell about this theory pair, we are asked to suppose that there are two scientists, Alice and Bob. Alice accepts the new theory whereas Bob accepts the old one. They are given a measurement device – call it the mystery box. As experienced scientists, they determine the features of the mystery box needed to calculate the expected value of performing a measurement on the system that Alice and Bob believe to be correctly described by T_{new} and T_{old}, respectively. When they compare their calculations, they observe two things. First, they reached the same result – they agree on what statistics to expect for the measurement outcomes. Second, each of them has worked in what the respective other would label 'momentum space'. When Alice asks Bob why he used the equation he used to model the system, he says: 'I just prefer to work in position space.' To this, Alice replies: 'But you didn't! *I* worked in position space. You worked in momentum space.' Bob of course rejoins: 'No, it's the other way around!'

What is going on? It is important to note that at this point in the story we might simply be dealing with a verbal dispute – it might be that Alice and Bob are speaking past each other because Alice means by 'position' what Bob means by 'momentum', and vice versa. This possibility will also become relevant later. Grimmer et al. however want to make it clear that there is supposed to be genuine disagreement between Alice and Bob. After all, the example is intended to be one of actually rival theories (or at least of theories for which a judgment of equivalence is non-trivial and must be based on a common core). So, the story continues in such a way that Alice and Bob are both adamant that they take the respective manifolds they are using to represent physical space, i.e., position space, not momentum space.

At this point however it is no longer clear that we are dealing with dual theories. In particular, it is not at all clear that the two theories are empirically equivalent. The basis on which Grimmer et al. claim that the two theories are empirically equivalent is that Alice and Bob calculated the same expected value for the experiment with the mystery box. At

the same time however their different theoretical commitments force Alice and Bob to disagree over the nature of the mystery box: Alice believes that it measures momentum whereas Bob believes that it measures position. But there's the rub: for plausibly there are independent ways to ascertain whether the mystery box measures position or momentum. After all, we have other devices of which we *know* what observable they measure. For position measurement, a simple screen will do. For momentum measurement, there are various spectroscopic methods. One can also use Bragg scattering or laser cooling etc. But this leads to a dilemma for Grimmer et al. To fully flesh out their story, they need to tell us what Alice and Bob predict will happen if we first perform the mystery measurement – which we may assume to be non-disturbing – and subsequently feed the system into a spectrometer. There are only two options: either Bob's predictions differ from Alice's. In that case their theories are empirically inequivalent. Or Bob insists that the spectrometer measures position, not momentum. But then it would seem that Bob is simply using the word 'position' to mean momentum. In that case, we are led back to the option we initially considered, that the disagreement between Alice and Bob and hence between T_{new} and T_{old} is merely verbal. Either way, the example Grimmer et al. give poses no threat to the common core account of equivalence.

But even if the example Grimmer et al. give does not cause problems for the common core account of equivalence, I take it that there is an example of a genuine duality lurking in the vicinity – we simply consider T_{new} as a theory of momentum space-time, not of spacetime. Much of what Grimmer et al. say about the (non-)existence of a common core in their example carries over to this case. So, it will have to be argued that the common core account of equivalence can accommodate this case.

I just suggested that if we can interpret Bob as meaning momentum by 'position' and position by 'momentum'; in other words, if we treat T_{old} as a theory of momentum space-time, then the appearance of disagreement between T_{old} and T_{new} disappears and the two theories are straightforwardly equivalent. But how straightforward is this really? As Grimmer et al. note, the usual way to see that the two theories are equivalent is to treat the two scalar fields as different representations of a state vector in a Hilbert space. But can this perspective on the two theories be considered a common core? Grimmer et al. argue that it cannot. The issue, they claim, is that in moving to the Hilbert space representation of

the system we are invoking structure – to wit, a Hilbert space – that is not found in either T_{old} or T_{new}.

However, I don't think this is right. When Grimmer et al. set up T_{old} and T_{new}, they characterize the kinematics of the theory in terms of the following condition on the scalar fields: "In some fixed global coordinate system, (t,x), each field must be square-integrable across x at each t-coordinate with unit norm." (Grimmer et al. [forthcoming], p. 8) But the insistence on the unit norm is gratuitous. If we abandon it, what are we left with if not a Hilbert space? So, if there is structure that is not common to T_{old} and T_{new}, it cannot be a Hilbert space. Nor can it be the preferred bases – each theory contains the position and momentum operators, but the preferred bases are just the eigenbases of these operators.

What else could it be then? One suggestion might be that T_{new} lacks a representation of momentum space and T_{old} lacks a representation of position space. After all, the base manifold is to be interpreted as space-time in T_{new} but as momentum space-time in T_{old}. But that is not convincing either. Position and momentum space appear as the spectra of the respective operators and so can be found in the theory after all (see Butterfield [2018]).

6 Horizontal and vertical reformulation

Having addressed most of the objections to motivationalism about dual theories, let me conclude this chapter by explaining in more detail why a coherent, well-motivated form of motivationalism remains even if we retreat from strict isomorphism as the standard of equivalence beyond which no further reformulation is required to the more inclusive standard of mathematical equivalence. One might worry that I was too cavalier in suggesting that this weakening leaves the core of the motivationalist outlook unaffected. After all, a basic motivationalist tenet is that we may only regard two theories as equivalent if we have a coherent account of the unified description we may take them to offer of physical reality. As we saw however, mathematically equivalent formulations of one and the same theory can at times be suggestive of starkly different ontologies and thus pictures of the world – recall the example of the Schrödinger and Heisenberg formulations of quantum mechanics for instance. I therefore want to show that the discussion of this

chapter really does leave us with a principled form of motivationalism, and one moreover that is strongly continuous with the standard motivationalism found in the literature.

To do so, let me first introduce a distinction between horizontal and vertical reformulation. Intuitively, horizontal reformulation transforms one structure into an equivalent one – we remain at the same level of structure, although that structure may be presented differently. Horizontal mutual reformulations are therefore mathematically equivalent. To borrow an example from Wallace, three-dimensional Euclidean space for instance can be defined as "a set coordinatized by a family of bijections into $\mathbb{R}^3$ [...], as a 3-dimensional affine space equipped with an inner product on its associated vector space, or as a 3-dimensional manifold diffeomorphic to $\mathbb{R}^3$ and equipped with a flat non-degenerate Riemannian metric" (Wallace [2022a], p. 353). Horizontal reformulation is what we find between theories that are interdefinable or Morita equivalent (Barrett and Halvorson [2016b]), for instance. Good examples from physics are General Relativity and the theory of Einstein Algebras on one hand (Rosenstock, Barrett and Weatherall [2015]) and, bracketing the historical wrinkles noted by Muller [1997], the Schrödinger and Heisenberg formulations of quantum mechanics on the other. We will see a further example from quantum mechanics below. In the case of vertical reformulation by contrast, we strip away structure, i.e., we descend from a more highly structured space to a less highly structured one. To illustrate this notion, any example of reduction or sophistication will do, such as for instance the move from Newtonian to Galilean spacetime.

Some clarifications are in order. First, the ordering of theories implicit in the talk of vertical and horizontal reformulation is merely partial. Even if T_1 and T_2 share a common core, i.e., even if there is a theory T^* that sits below both T_1 and T_2, T_1 and T_2 can be incomparable according to the implied ordering. In other words, it might be that T_1 and T_2 are neither horizontally nor vertically related. Secondly, as they're in general not mathematically equivalent, the reduction and sophistication of a theory will in most cases not be horizontally related, even though it may be tempting to think so.

Put in terms of horizontal and vertical reformulation, the motivationalism I defend then is this: we can interpret theories as equivalent whenever they have a common reformulation, be it a horizontal or a vertical reformulation. In case it is a vertical reformulation, the bottom theory is a core theory of the top theory or theories. Some might worry that being so

liberal about treating horizontal reformulations as equivalent dilutes motivationalism too much. Since there are many cases of mathematically equivalent theories that strike us as ontologically inequivalent, it would seem that I am happy to declare even theories with incompatible ontologies to be equivalent. That however sounds like we've completely abandoned the core motivationalist view that we can only interpret theories as equivalent if we have a coherent understanding of how they could involve the same ontology.

To dispel this concern, let me point out that we already understand quite well how to view mathematically equivalent theories as ontologically equivalent. Broadly, the crucial insight is that equivalent formulations of the same theory can always be seen as involving the same ontology as long as we treat them as differing in how ontologically perspicuous they are. When I speak of ontological perspicuity here, I propose that we understand this notion along the lines of Jacobs' [2022a] definition of intrinsicality. In other words, I will take a theory formulation to be perspicuous just in case it is formulated in terms of a "set of relations, functions and operators which explicitly represent the world's physical structure" (Jacobs [2022a], p. 15), although I prefer to speak of direct, rather than explicit representation.

On this understanding, horizontally related reformulations, while having the same content, can exhibit wildly different degrees of ontological perspicuity. This often creates the illusion that they involve incompatible ontologies when in fact they do not. For an example taken from the literature on paraphrase (Keller [2017]), contrast the following two sentences

(1) The average person has 2.4 children
(2) The sum of the numbers of children of every person divided by the total number of people is 2.4

While (1) and (2) are equivalent and related by a horizontal reformulation, (2) is more perspicuous than (1) because each singular term in (2) directly refers to an object, whereas the same is not true of (1) – there is no such thing as 'the average person' after all. Horizontal reformulation then aims at ontological perspicuity, whereas vertical reformulation aims at the elimination of unwanted structure. It seems to me now that the distinction between perspicuous and imperspicuous formulations of theories in physics

can be understood precisely along these lines. To take a well-worn example, consider the ordinary formulation of n-particle classical mechanics in terms of n points in three-dimensional Euclidean space, and as its counterpart the reformulation of that theory set on phase space. This is another instance of horizontal reformulation – as Wallace and Timpson point out, the two formulations are "naturally isomorphic" (Wallace and Timpson [2012], p. 701). And once again, there is a stark contrast in perspicuity owing to the fact that the building blocks of the first theory – three-dimensional Euclidean space and n points in that space – directly represent physical entities, whereas phase space and a privileged point therein do not have physical correlates. Rather, the phase space theory encodes information about three-dimensional space and the particles therein indirectly. Nonetheless, the theories of course involve exactly the same ontology.

Of course, there may not always be a way to tell which of two mathematically equivalent, i.e., horizontally related theories is more ontologically perspicuous than the other. This notwithstanding, the insight remains: mathematically equivalent theories can always be understood as involving the same ontology even when on the surface they seem incompatible. This now allows us to retain the motivationalist idea that judgments of equivalence must be supported by an account of the ontology common to the putatively equivalent theories. It is just that any one of the available mathematical formulations of the equivalent theories' common core must be viewed as giving us this account, however obliquely and imperspicuously it may do so.

Let me illustrate the point that mathematically equivalent theories can be interpreted as having the same ontology further. To put it in a slogan, what I am suggesting is that ontological change is vertical, not horizontal. This is important, as it goes against the lessons one might be tempted to draw from certain instances of horizontal reformulation, viz. that mathematically equivalent theories can posit incompatible ontologies. That ontological change is vertical, not horizontal is a point very well made by Wilson [1981], who furthermore identifies a range of nicely illustrative examples from the history of mathematics. The first of his examples make it clear how instances of horizontal reformulation can mislead one into thinking that a change in ontology has occurred. Take first the standard theory of the complex numbers and compare it to a definition of the complex number structure as $\mathbb{R}^2$ equipped with appropriately defined multiplication and conjugation operations. Secondly, take the standard definition of the real numbers as a

Dedekind complete ordered field and compare it to Dedekind's construction of the reals as cuts of the rational numbers. These reformulations are often hailed as paradigms of 'ontological reduction': a chain of reformulations that reduces various number systems to constructions from the natural numbers. But that of course invites one to think of these reformulations as effecting a change in ontology: the first eliminates the complex numbers as *sui generis* entities, just like the second eliminates the real numbers, the thought might go. As Wilson emphasizes, it would be a mistake to draw this conclusion. In each of the examples, the theories involved define exactly the same class of structures, since they are mathematically equivalent. The ontologies of the theory pairs are therefore in each case one and the same. *Vertical* reformulation on the other hand does lead to ontological change: here, Wilson gives as an example the Cauchy-Weierstrass theory of limits, viewed as a successor to the theory of infinitesimals. The new theory genuinely does away with the infinitesimal quantities countenanced by the old and so does involve a shift in ontology.

To return closer to home, one can also illustrate Wilson's point with examples from physics. A particularly important such example is discussed in Bokulich [2020]. As it turns out, there is a reformulation of quantum mechanics whose mathematical formalism does not employ wave functions. That theory is called Lagrangian quantum hydrodynamics.[97] In quantum hydrodynamics, the quantum state is given by a displacement function $q(a,t)$ for a continuum of particles flowing through three-dimensional space along smooth trajectories, the congruence of which is captured by $q(a,t)$. In a slogan, this gives us "quantum mechanics without wavefunctions" (Bokulich [2020], p. 204). But we must be careful not to misinterpret the import of this slogan. As Bokulich emphasizes, LQH is an exact reformulation of standard quantum mechanics, fully equivalent to the wave function formulation. In other words, we have an instance of horizontal reformulation. Thus, however tempting it may be, it would be a mistake to think that LQH "eliminates" the wave function in the sense of eschewing structure present in the standard formulation of quantum mechanics.[98] The correct thing to say is a different one: contra wave function realists (see e.g. Albert [2013], Ney [2021]), quantum hydrodynamics shows that quantum theory does not force us to *reify* the wave function and the configuration space on which it is defined, by which I mean to think of configuration space as a faithful representation of

[97] See Holland [2005], [2013] for an exposition of LQH

[98] To be clear, I am not suggesting that Bokulich makes this mistake – her discussion of the example is very rich and subtle and strikes me as convincing. I am merely pointing out that the example invites a misinterpretation along the lines sketched.

physical space and of the wave function as directly representing an entity in that space.[99] In other words, quantum hydrodynamics does not alter the mathematical structure of quantum theory. It therefore possesses exactly the same ontology as the wavefunction formulation. However, quantum hydrodynamics cautions against facile (or at least hasty) attempts at determining the ontology of quantum mechanics via a principle according to which the wavefunction formalism must be interpreted as an intrinsic formulation of quantum mechanics, the mathematical objects of which directly and faithfully represent structure in the world. In Wallace's [2022a] terminology, quantum hydrodynamics shows us that the wavefunction formulation is not the only game in town and that therefore its predicate precisification need not be regarded as a definitive guide to the ontology of quantum theory. The important point for our purposes is this: regardless of what the correct ontology of quantum mechanics will turn out to be in the end, it will be shared by the various horizontal reformulations of the theory. In particular, since quantum hydrodynamics and the wavefunction formalism are mathematically equivalent, they can always be interpreted as involving the same ontology. Once again, horizontal reformulation does not bring about ontological change.

To summarize the most important points made in this chapter, I have argued that for it to be permissible to interpret two theories as equivalent, it must be established that they can be viewed as representations of the same core formalism. However, this does not invariably amount to giving a coherent account of their shared ontology. The common core may raise its own interpretive difficulties. So, *pace* some motivationalists, we can in principle be warranted to interpret two dual theories as equivalent without fully understanding the ontology that grounds their equivalence. This insight allowed us to formulate a more moderate form of motivationalism which, as I have argued, can accommodate one of the central objections to motivationalism, i.e. the 'equivalence without interpretation' objection. It also gave us the liberty to weaken our standard for when two theories may be interpreted as equivalent without reformulation from strict isomorphism to mathematical equivalence. This in turn gave us the resources to defuse the 'equivalence without common core objection' Crucially, despite the concessions we made in responding to these objections, a coherent, well-defined, and substantial motivationalist thesis survives: we need a common core to interpret dual theories as equivalent whenever they are mathematically inequivalent. In the next chapter, I will argue that this moderate

[99] I am following Bokulich in drawing this conclusion

motivationalism can handle the third of the objections to motivationalism we identified, i.e. the 'scholastic metaphysics' objection.

7 The Problem of Conceptual Relativity

1 Introduction

In the previous chapter, we saw that motivationalism is in tension with the phenomenon Putnam [1977, 1980, 1983, 1987, 2001] calls 'conceptual relativity'. This chapter will attempt a fuller examination of this tension. In a slogan, recall, the phenomenon of conceptual relativity involves "equivalent, but incompatible" descriptions of the world. By this, Putnam means equally good perspectives on, or conceptualizations of the world, which however cannot be adopted simultaneously (see e.g. Putnam [2001]). In this context, he often speaks of incompatible conceptual schemes. The idea of "equivalent but incompatible" theories of course goes against the spirit of motivationalism. After all, whenever two theories are dual but seemingly incompatible, motivationalists maintain that we ought to reconcile them by seeking a coherent account of the ontology that underlies them. The phenomenon of conceptual relativity calls into question whether such a coherent account is always to be had. So, let us speak of the *problem of conceptual relativity* for motivationalism.

The problem of conceptual relativity comes in a mild form and a more severe form. In its mild form, it is the problem, pointed out *inter alia* by Wallace [2022a], that mathematically equivalent descriptions can be suggestive of different ontological pictures – in Wallace's [2022a] terminology, different formulations of the same theory often invite different "predicate precisifications", i.e. formulations of the theory in different languages that seem to speak of different kinds of objects. Those who believe in conceptual relativity maintain that each of these different predicate precisifications is equally legitimate. The thesis that one of them is privileged in the sense of being especially perspicuous, or especially congruous with the fundamental structure of reality, is called metaphysical realism:

> As I explained metaphysical realism, what it came to was precisely the denial of conceptual relativity. My metaphysical realist believed that a given thing or system of things can be described in exactly one way if the description is complete and correct, and that way is supposed to fix exactly one ontology and one ideology in Quines sense of those words, that is, exactly one domain of individuals and one domain of predicates of those individuals
>
> (Putnam [2012], p. 62)

That different definitions of one and the same class of structures often invite different predicate precisifications puts pressure on metaphysical realism. This however is not in itself a problem for motivationalism. To be sure, many motivationalists maintain that judgments of equivalence must be underwritten by a clear picture of the ontology shared by the theories in question, and this view has a clear affinity to metaphysical realism. And perhaps many motivationalists who hold this view really share a background commitment to metaphysical realism. But as we noted, these background beliefs are separable from a well-defined core motivationalist thesis, according to which judgments of equivalence for models or theories are justified *ab initio* just in case the relevant models or theories are mathematically equivalent, and else must be justified by demonstrating that the two theories share a common core. This core thesis was a weakening of the standard motivationalist view according to which the standard of evident equivalence is strict isomorphism of models, not the looser standard of mathematical equivalence. This more moderate motivationalism emphasizes the importance of reformulation, and relegates the kind of informal accounts of the ontology shared by putatively equivalent theories that one might want to give in the form of a predicate precisification. For a motivationalist of this stripe, it is a matter of comparative unimportance whether the common core theory grounding the equivalence of two theories admits different mathematically equivalent characterizations and hence perhaps different predicate precisifications or "ontological pictures". That being said, motivationalists are free to maintain of course that among a theory's different predicate precisifications, one is preferred on the grounds of being especially ontologically perspicuous or joint-carving. But they need not do so. In any case, in its mild form the problem of conceptual relativity demands nothing but a slight refinement of the motivationalist's standpoint.

The problem of conceptual relativity can take a more severe form however. In a nutshell, this happens when we have (a) two mathematically *in*equivalent theories with incompatible ontologies that (b) by all accounts are interchangeable as descriptions of the world and (c) do not seem to possess a common core. Call such theory pairs cases of *strong conceptual relativity*. Here are some putative examples of strong conceptual relativity.

Example 1: Convergent spheres

The first example is taken from geometry. While it is common to take points in space as the primitive subject matter of geometry, Putnam points out that one can equally well have an

axiomatization of geometry whose basic entities are spheres (see Putnam [1989]). Points appear in such a theory as constructions from spheres, more precisely as sequences of convergent spheres. The two theories seem to involve incompatible ontologies: the first countenances extensionless entities whereas in the second, all objects are spatially extended. Indeed, Putnam emphasizes that there is a precise mathematical sense in which the theories are incompatible:

> if we simply conjoin them without taking account of the different ways the terms are used, we get a contradiction at once. [...] E.g., if we conjoin a theory in which points are sets of convergent spheres [...] with a theory in which spheres are sets of point [sic] or mereological sums of points, and we assume the Axiom of Foundation of set theory we get an immediate contradiction.
>
> (Putnam [2001], p. 436)

Nonetheless, the theories seem equivalent. As Putnam points out, "[i]t is known since Principia Mathematica at least that we can identify points with sets of convergent spheres and all geometric facts will be correctly represented." (Putnam [1989], p. 112)

Example 2: Points and lines

Just like the first example, this example, too, is taken from geometry. As it turns out, points can be viewed as constructions not only from spheres but also from lines. We can therefore axiomatize geometrical theories in a way that treats lines as primitive entities (see Putnam [1978]; Barrett and Halvorson [2017]). Once again, we are faced with apparently perfectly equivalent theories of geometry that nonetheless involve seemingly incompatible ontologies. This example has been discussed at length by Barrett and Halvorson [2017], who argue that in this case, the ontologies can be reconciled via a construction called a *Morita extension*. We will discuss below whether their argument succeeds in undermining the status of the example as an instance of strong conceptual relativity.

The next two examples are mentioned by Sider (Sider [2020], Ch. 5) and luckily are cases we have already considered in previous chapters.

Example 3: NGT in g vs NGT in kg

Take a version of NGT in which objects are assigned a mass in grams and compare it to a version of NGT in which masses are given in units of kilograms. These theories may not strike one as incompatible, but they are mathematically inequivalent because each involves

a different distinguished mapping of particles into the reals. Metaphysically, the disagreement between the two is over which mass is the true unit weight.

Example 4: The geometry of spacetime
There is a tradition of conventionalism about the geometry of spacetime (see e.g. Reichenbach [1958]). According to such a view, different ascriptions of a geometry to spacetime can be "equivalent but incompatible" in the sense that they amount to different choices of convention. No such choice is privileged. Most recently, the geometric trinity (see Chapter 6) has been adduced in support of a conventionalist account of the geometry of spacetime (Dürr and Read [2023]).

At least initially, these examples invite an anti-motivationalist attitude. One is tempted to conclude that even though it may be hopeless to try and construct a common core for these theories, clearly, they are equivalent. But of course, in Examples 3 and 4 it turns out that a common core can be found after all. A theory of mass values in terms of additive extensive structures can reconcile the apparently incompatible versions of NGT (see Chapter 4, Sider [2020], Dewar [2024], Jacobs [2022]). And in the case of the geometric trinity, we saw above that GR constitutes a common core theory. We see immediately therefore that the common core account of equivalence gives motivationalists resources for handling alleged examples of conceptual relativity. As I will argue below, Examples 1 and 2 can also be accounted for by common core constructions, although the case is philosophically more subtle. What I aim to show with this is that the common core account of equivalence gives motivationalist a powerful tool for dealing with examples of strong conceptual relativity. But to fully develop the argument, let us first look at the idea of conceptual relativity in detail.

2 Conceptual Relativity

Before we investigate how motivationalists ought to deal with the problem of conceptual relativity, let us go into more detail on what this phenomenon involves. In a slogan, as we said, conceptual relativity is the existence of "equivalent, but incompatible" descriptions of the world – pairs of theories that by all accounts do an equally good job of describing how things are, but which do so in terms of different, incompatible ontologies. One of Putnam's preferred examples is that of mereological nihilism and mereological universalism (see

Putnam [1987]). Mereological nihilism is the view that all objects are simple, without proper parts. Universalism by contrast maintains that any number of objects fuse to form a composite object of which they are parts. These theories are clearly incompatible, and yet Putnam maintains that they yield equally adequate descriptions of the world. After all, he claims, any state of affairs that can be described in the language of the universalist can equally well be described in the language of the nihilist, and vice versa. Indeed, it is a test for when two incompatible conceptual schemes are equivalent that every state of affairs describable in one scheme and possible according to that scheme is also describable in the other and possible according to the other, and vice versa (Putnam [1987]).

One helpful way to think about conceptual relativity appeals to the notion of a conceptual scheme or "optional language" (see Case [2001]). In brief, a conceptual scheme is a way of conceptualizing the world. For instance, mereological nihilism and universalism are different conceptual schemes for Putnam. Since the world can only be described or represented by way of a conceptual scheme, one must choose such a scheme in order to express one's theories. Hence the notion of conceptual relativity, i.e. relativity to a conceptual scheme. Crucially now, there will in general be multiple schemes to choose from. As the example of mereological nihilism and universalism illustrates, these schemes are not compatible. Nonetheless, there is no "right" or "wrong" scheme. The choice between schemes is guided by pragmatic considerations alone. It is a matter of convention. In much the same way that driving on either the left or the right side of the road are equally good but incompatible conventions, different conceptual schemes offer alternative ways to describe the world which however cannot be combined.

That a choice between schemes is a choice between conventions is a point worth emphasizing: by way of diagnosis, Putnam maintains that the phenomenon of conceptual relativity results precisely from the inextricability of fact and convention.[100] Thus, Putnam states:

> The doctrine of conceptual relativity, in brief, is that while there is an aspect of conventionality and an aspect of fact in everything we say that is true, we fall into hopeless philosophical error if we commit a "fallacy of division" and conclude that there must be a part of the truth that is the "conventional part" and a part that is the "factual part" (Putnam [1990], p. x, cit. Case [2001])

[100] For more on the interwovenness of fact and convention see Quine [1954]

A fortiori, if there is no way to separate fact and convention, there will not be a way to capture only the factual content of a theory, entirely free of conventional aspects. Insofar as such a convention-free, "bare" representation of a theory's content is the regulative ideal behind the motivationalist's search for common cores, Putnam's view is that such representations are in general unattainable due to the necessary, arguably transcendental, role of convention in representation, which in turn manifests in the phenomenon of conceptual relativity (Putnam [2001]). It is in this sense that conceptual relativity poses a threat to the motivationalist position.

The sentiment that conventionality plays an ineliminable role in scientific representation is echoed by those Sider [2020] calls *quotienters*. Quotienters, much like interpretationalists, maintain that it can be legitimate to regard mathematically inequivalent models as differing only in conventional aspects and hence as equivalent even without providing a convention-free common core. Their view is rooted in the conviction that

> [t]here may be no way to say what is "really" going on; maybe every good model has artifacts. It's ok to just say: this model does a good job of representing the phenomenon, but certain features of the model are artifacts. Moreover, for any model, we can say which features of the model are genuinely representational and which are artifacts. There is no need to provide some privileged, artifact-free description from which we can recover this information. (Sider [2020], p. 153)

Two points are worth highlighting about this quote. First, note that Putnam is even more radical than Sider's quotienter. Whereas the quotienter thinks that we can separate out the conventional from the representational, Putnam denies this. He sees the two as inextricably linked. Secondly, note the emphasis in the quote on the anti-motivationalist implications of the quotienter's view: we can be warranted in treating two mathematically inequivalent models as equivalent even when we don't have a convention- and artifact-free account of the reality they both represent which would allow us to understand how they are both equally representationally successful. Reformulation is seen as unnecessary because it is unattainable. An artifact-free common core is an elusive, unrealizable ideal.

Defenders of conceptual relativity rest their view on intuitively powerful examples. The question we are facing therefore is how motivationalists should deal with the problem these examples create. In asking this question, it is important to emphasize, as I did in the last chapter, that motivationalists can accommodate Putnam's insights to a considerable extent. Adherents of the moderate motivationalism I have been defending will regard the

existence of equivalent theories which are nonetheless suggestive of different ontologies with equanimity, provided that the theories are mathematically equivalent. Such examples of "mild" conceptual relativity are perfectly compatible with moderate motivationalism. Motivationalists can even grant Putnam (and Wallace [2022a]) that in such cases there is no sense in which one of the equivalent formulations of the theory is more perspicuous or natural than the others (although they need not do so).

Where motivationalists will have to part ways with Putnam however is when it comes to putative examples of *strong* conceptual relativity, i.e. examples that involve mathematically *in*equivalent theories. In such cases, Putnam's [2001] attitude is that of the quotienter: we can recognize that the theories in question are inequivalent in the absence of a common core theory that reconciles them, and indeed the hope to find such a common core may be in vain, given the inextricability of fact and convention. And of course it is precisely this inevitable "convention-ladenness" of theories that should make one expect to find instances of conceptual relativity involving mathematically inequivalent theories. After all, why should different conceptual schemes invariably employ equivalent mathematical representations?

One thing motivationalists ought to point out about such cases is that the judgments of equivalence Putnam and his fellow quotienters are prepared to make are bound to be on less secure footing than they might think. Motivationalists are right to worry that in cases of mathematically inequivalent theories, we simply cannot know that all the factual content is the same, even if we incline towards a judgment of equivalence. Often, there turn out to be salient differences between the theories that we have overlooked. And indeed, many of Putnam's alleged examples of conceptual relativity falter under closer examination. Frequently, the theories Putnam adduces as examples of conceptual relativity turn out not to be equally apt to describe the world after all. Putnam [1989] treats the case of vector potential and E/B-field electrodynamics as an instance of conceptual relativity, even though there are significant differences between the two theories. And even those cases that have to many seemed as though they must involve interchangeable theories often do not withstand scrutiny. Take for instance Putnam's [1987] example of mereological nihilism and mereological universalism, for many a paradigm of merely verbal disagreement and idle metaphysical speculation. Contrary to what Putnam claims, the two theories are not equivalent. There are robust modal differences between them. According to mereological

nihilism, necessarily, whenever something exists, some simples exist. Not so according to mereological universalism, which admits non-empty possible worlds without simples – gunky worlds, say, where every object has proper parts.

The lesson here – and it is a motivationalist lesson – is not to declare theories to be equivalent prematurely, just because it is not immediately obvious what hinges on their differences. There is a tendency to write off disagreement between theories as merely verbal whenever it seems difficult to adjudicate between them or the debate over which is correct threatens to get lost in the thickets of speculative metaphysics. Of course, it is salutary to take a step back now and then and ask whether a disagreement concerns anything of substance or whether its participants are talking past each other. But the tendency just described can make one too cavalier and blind to significant, albeit perhaps subtle, differences between apparently interchangeable theories. I suspect this tendency is at least part of what makes Putnam's discussion of conceptual relativity seem like such a powerful threat to motivationalism.

That being said, the problem of strong conceptual relativity cannot be fully defused by (rightly) urging caution when it comes to judgments of equivalence for mathematically inequivalent theories. There remain cases of at least apparently mathematically inequivalent theories which just seem equivalent. What is worse, many of these cases look as though the common core approach cannot be brought to bear on them to achieve reconciliation. In such cases, the quotienter's attitude becomes very tempting.

In what follows, I will consider a number of such cases, and in particular I will consider attempts at construing the theories in question as mathematically equivalent after all. I will argue that while not all strategies for doing so succeed, the cases in question can be handled by the common core method. This shows that the motivationalist has the resources to deal with these putative examples of strong conceptual relativity. Of course, this does not rule out that there might be genuine examples of strong conceptual relativity after all. But it shifts the burden onto the quotienter to provide us with such examples. Moreover, the rarer and the more tenuous such examples, the more convincing the motivationalist's insistence that in such cases, we simply cannot know that the theories are equivalent unless we find a common core.

3 The spectre of strong conceptual relativity

We will look at ways motivationalists might attempt to undermine putative instances of "strong" conceptual relativity. Specifically, we will consider two arguments to the effect that certain theory pairs that *prima facie* seem to be instances of strong conceptual relativity can be viewed as having the same underlying ontology after all. Each of the two arguments appeals to certain formal equivalence results for the theories in question to argue that despite appearances, the theories are compatible and indeed fully equivalent.

3.1 The Reinterpretation strategy

The first strategy for reconciling theory pairs that putatively illustrate the phenomenon of conceptual relativity is one discussed *inter alia* by Putnam [1978] and Wilson [1981]. With this strategy, the aim is to show that two theories can be understood as compatible if we interpret some of the terms ostensibly shared by them as diverging in meaning. The strategy at hand therefore falls under an approach to resolving disputes between theories one might call the 'meaning variance' approach: it is shown that while proponents of the two theories use the same words, they attach different meanings to them. Once that is realized, the dispute can be recognized as a merely verbal one and is thereby resolved. This treatment is applicable to a wide range of putative examples of conceptual relativity. For the sake of concreteness, I will first present it using one particular example. But I will subsequently show how the strategy generalizes to other cases. I will argue that while the strategy is not successful, motivationalists have the resources to account for the equivalence of the theories on which it is brought to bear.

Schematically, the idea is to first construe the putatively incompatible theories – call them T_1 and T_2 – against a set theoretic background. For a concrete example, taken from Wilson [1981], consider the theory pair of standard set theory, ZFC, and a set theory constructed over the natural numbers as urelements – call that second theory ZFCN. ZFCN will add to the axioms of ZFC the Peano axioms, say, to govern the naturals that are regarded as urelements. No technical details beyond this are required for the discussion. As Wilson notes, at first glance, these theories have the air of incompatibility. After all, according to the first theory, there are only pure sets, whereas the second theory also countenances

non-set urelements. Despite these apparently incompatible ontologies however, descriptively, the theories seem equivalent. A textbook case of conceptual relativity then, it seems.

The next step is therefore to argue that despite appearances, the theories have the same ontology after all. This argument proceeds on the basis that the theories can be shown to be interdefinable if we allow ourselves to give some of the vocabulary shared between the two theories a deviant interpretation in one of the theories (Wilson [1981]). In this case, as is typical, this involves reconstruing the '$\in$' symbol as it occurs in ZFCN as standing not for what from the perspective of ZFC is the standard elementhood relation $\in$ but for a different relation $\in^*$. The vocabulary not shared between the two theories, i.e. the relation and function symbols of PA must be defined as well of course, but this can be done since ZFC interprets PA. The claim is then that due to their interdefinability the two theories are equivalent and have one and the same ontology – the set theoretic universe –; it is just that '$\in$' as it occurs in ZFCN has a different meaning from the '$\in$' of ZFC. Whether the argument is convincing or not remains to be seen of course.

This strategy carries over in an obvious way to other putative examples of conceptual relativity. First, consider Putnam's example of thinking of points (a) as primitive or (b) as derivative of sequences of nested spheres, those spheres in turn being thought of as primitive. We can apply the strategy by first associating each of the viewpoints with an extension of ZFC. Call the resulting theories ZFCP and ZFCS, respectively. ZFCP is a set theory over a space of urelement points – say, three-dimensional Euclidean space. ZFCS has for its urelements a class of spheres, described by an axiomatic theory S which captures all the definable properties of spheres in 3D Euclidean space. One then shows that ZFCP and ZFCS are interdefinable provided a reinterpretation of the '$\in$' symbol is admitted. As before, terms not shared between the theories must be defined, which can be done since spheres are definable as sets of points and points are definable as limits of sequences of spheres.

Secondly, consider the example discussed by Barrett and Halvorson [2016], i.e. the example of certain geometries axiomatized alternatively in a way that treats points as primitive and a way that treats lines as primitive. By now, the procedure ought to be clear: We construe the first theory as a theory ZFCP, with a geometry P axiomatized in terms of

urelement points against a background set theory ZFC. We do the same for an axiomatization in terms of lines L of the relevant geometry, to obtain a theory ZFCL. Finally, we show that ZFCP and ZFCL are interdefinable under a deviant interpretation of '$\in$'.[101]

But how convincing is this argument? The technical interdefinability results are unimpeachable of course. But the argument hinges also on a semantic thesis. Crucial to its success is the claim that in the examples at hand, the meaning of '$\in$' varies from one theory to the other. This claim is doubtful however. To focus on just one of the three examples (without loss of generality), take the theory pair ZFC and ZFCN. We can imagine two mathematicians, Zach and Natalie, who disagree over the correct ontology for mathematics. Zach defends ZFC whereas Natalie advocates for ZFCN. Zach and Natalie, let us suppose, are both native speakers of English and expert set theorists who obtained their PhDs under the same supervisor. If they don't know the meaning of '$\in$', nobody does. Let's say moreover that they have co-authored papers before and therefore have endorsed the very same set theoretic claims in the past. It seems highly plausible that when Zach utters 'ZFC is the correct foundation of mathematics.' and Natalie utters 'No, ZFCN is.', they are using the language of set theory under the exact same interpretation; they are not talking past each other but simply disagreeing with each other.[102] Nor is there anything odd about the idea that Zach and Natalie should be arguing about the correct foundational theory for mathematics. Set theorists frequently disagree over new candidate axioms to extend ZFC such as large cardinal axioms or the continuum hypothesis, and many of them understand their dispute in the most straightforward way, as about what is true of the sets. An analogy might help: think of Zach and Natalie as two distinct models of vector potential electrodynamics within the same gauge orbit. All the vocabulary they use has the exact same meaning, and they describe the same target system. It's just that they make incompatible claims about that system.

Those who take the theories ZFC and ZFCN to be an instance of strong conceptual relativity can therefore insist that their interdefinability modulo reinterpretation of '$\in$' fails to establish that the theory pairs in question share an ontology, because the reinterpretation of '$\in$' is illicit. In each theory, the intended interpretation of '$\in$' is the same, namely set

[101] I don't know of an explicit demonstration of how this is done, but Putnam [1977] claims a proof exists.

[102] Arguments of this kind for the claim that there is no deviation in meaning between the two theories are familiar from e.g. Williamson [2007], Ch. 5

theoretic elementhood. That the theory pair becomes interdefinable if one assigns to '$\in$' a new, unintended meaning therefore no more shows that they are equivalent than the fact that 'The zoo owns three lions' and 'The zoo owns three tigers' are interdefinable if one interprets 'lions' to mean tigers shows that these two theories are equivalent – 'lions' refers to lions, not tigers, and '$\in$' refers to $\in$, not $\in^*$, and that is that.[103]

One might worry therefore that in this case, and indeed in each of the three examples, we face pairs of mathematically inequivalent theories that nonetheless are equally eligible as candidate descriptions of the world. After all, whether one describes points as constructions from lines or lines as constructions from points is surely a matter of convention. Analogous points apply to the other examples. This puts pressure on motivationalists to treat the theory pairs in question as equivalent despite the fact that this verdict cannot be grounded in a formal equivalence. We face the problem of conceptual relativity in its strong form. But all is not lost. While the interdefinability argument for the equivalence of the theories at hand is doubtful, motivationalists can employ the machinery of common cores to interpret them as equivalent after all. The crucial observation in each of the three cases is that the structure on the spaces of urelements is definable in ZFC. In the case of ZFCN, we find the structure of an ω-sequence to be definable and instantiated in the pure sets. Obviously, the geometric spaces that provide the urelements in the theories ZFCP, ZFCL, and ZFCS are likewise definable within ZFC. In each of the theories, we can therefore view the urelements as surplus structure. This gives us as the common core of all the theories ZFC, the theory of the pure sets. This in turn allows us to say that Zach and Natalie are genuinely disagreeing. At the same time, there is an interpretation of Natalie's theory on which it becomes equivalent to Zach's theory.

In this section, we have seen one way in which motivationalists may try to defuse the threat of strong conceptual relativity. In outline, the idea is to construe a whole range of putative examples of strong conceptual relativity against a set-theoretic background and to argue that pure ZFC provides a common core for all of them. One advantage of this approach is of course its sweeping generality. Moreover, we are relying on the machinery of common cores, which makes the approach especially congenial to motivationalists. At the same time, the approach is not free of problems: in employing a background theory as strong as ZFC, we are committing to an ontology that has struck many as inordinate and potentially

[103] Sklar [1982] makes a similar point

beset with insurmountable conceptual problems. More generally, one might worry that there is something illicit about construing the putative examples of conceptual relativity against an extraneous background of set theory. It is arguably preferable to regiment the theories involved in a pure, intrinsic manner. Doing so however deprives us of ZFC as an available common core. Nonetheless, we will be able to tackle many alleged examples of strong conceptual relativity with the common core approach, even if we formalize them intrinsically. In the next section, we turn to showing this.

3.2 Morita equivalence

A different approach to dealing with the problem of conceptual relativity has been put forth by Barrett and Halvorson [2017]. In general, the idea is to formalize any pair of rival theories Putnam claims are an illustration of conceptual relativity in a language of many-sorted language and to then show that they satisfy a standard of equivalence for many-sorted theories called *Morita equivalence*. Barrett and Halvorson have shown in detail how to bring this strategy to bear on the example of geometries axiomatized alternatively in a language that has a primitive term for points or a language with a primitive term for lines. Thus, they show that any geometry meeting certain natural conditions can be axiomatized as a theory T_P in a language $\mathcal{L}_P$ with only a sort for points, and in a Morita equivalent way as a theory T_L in a language $\mathcal{L}_L$ with only a sort for lines. They claim that this shows the two theories to be equivalent simpliciter. *A fortiori*, they claim it shows that the theories have compatible, indeed identical, ontologies. Our task is to examine whether their argument is convincing, and whether the approach they propose is one that can be adopted by someone with the background commitments of a motivationalist.

First, let us say a little more about the technical background to their discussion. I will presuppose familiarity with the syntax and semantics of many-sorted logic. For an introduction to this topic, the reader is referred to Barrett and Halvorson [2016]. To define what it is for two many-sorted theories to be Morita equivalent, we must first introduce the notion of a Morita extension of a many-sorted theory T. Informally, a Morita extension of a theory T is what one gets if one constructs new sorts from the sorts already present in T via a number of admissible constructions. More precisely, a Morita extension T^+ of a theory T is of the following form

$$T^+ = T \cup \{\delta_\sigma \mid \sigma \in \Sigma(T^+)\backslash\Sigma(T)\} \cup \{\delta_\Pi \mid \Pi \in \Sigma(T^+)\backslash\Sigma(T)\} \tag{7.1}$$

where $\Sigma(T^+)$ and $\Sigma(T)$ are the signatures of T^+ and T, respectively and for each new sort symbol σ and new predicate symbol Π, δ_σ and δ_Π are definitional formulae for σ and Π, respectively. For new predicate symbols Π, δ_Π takes the form of a simple explicit definition, i.e.

$$\delta_\Pi = \forall_{\sigma_{i_1}} x_1 \ldots \forall_{\sigma_{i_n}} x_n \big(\Pi x_1 \ldots x_n \leftrightarrow \varphi(x_1, \ldots x_n)\big) \tag{7.2}$$

Matters are slightly more complicated for new sorts σ. The new sorts a Morita extension introduces are to be thought of as constructed from the theory's original stock of sorts according to a circumscribed list of methods of construction. They fall into three categories: product sorts, subsorts, and quotient sorts.[104] Intuitively, a product sort corresponds to taking the Cartesian product of two sorts, a subsort to taking a subset, and a quotient sort to quotienting a sort by an equivalence relation, i.e. to introducing a new sort the elements of which are equivalence classes of elements of the old sort.

In the case of σ being a product of sorts σ_1 and σ_2, the definitional formula δ_σ is of the form

$$\forall_{\sigma_1} x \forall_{\sigma_2} y \exists_\sigma^{=1} z(\pi_1(z) = x \wedge \pi_2(z) = y) \tag{7.3}$$

Intuitively, the functions π_1 and π_2 are to be thought of as projections of elements of the Cartesian product of the sorts σ_1 and σ_2 onto their first and second component, respectively.

Where σ is a subsort of a sort σ_1, the definitional formula looks like this:

$$\forall_{\sigma_1} x\big(\phi(x) \leftrightarrow \exists_\sigma z(h(z) = x)\big) \wedge \forall_\sigma y \forall_\sigma z(h(y) = h(z) \rightarrow y = z) \tag{7.4}$$

Here, ϕ defines a subset of σ_1 – precisely the σ_1s that are ϕ, and h is a bijection between the ϕs in σ_1 and their isomorphic copy, i.e. σ. Note that it is only admissible to extend a theory T by introducing a subsort σ of a sort σ_1 if T proves that σ_1 is non-empty. We call this an admissibility condition for the introduction of σ.

Finally, where σ is a quotient sort of a sort σ_1, we have the following definitional formula:

$$\forall_{\sigma_1} x \forall_{\sigma_1} y\big(\epsilon(x) = \epsilon(y) \leftrightarrow \phi(x, y)\big) \wedge \forall_\sigma z \exists_{\sigma_1} x\, \epsilon(x) = z \tag{7.5}$$

[104] Barrett and Halvorson's definition of Morita extensions also allows for a fourth kind of construction, coproduct sorts. But coproduct sorts lead to trouble (see McEldowney [2020]) and are not needed for the constructions relevant to this chapter, so I omit them.

The idea here is that we are quotienting σ_1 by an equivalence relation ϕ. Consequently, we require that phi define an equivalence relation on sigma_1 modulo the background theory T. ϵ is then a function that maps every element of σ_1 to its equivalence class in σ.

Equipped with the concept of a Morita extension, we can now define Morita equivalence. Thus, two theories T_1 and T_2 are Morita equivalent just in case there are sequences of theories $T_1^{(1)}, \ldots T_1^{(n)}$ and $T_2^{(1)}, \ldots T_2^{(m)}$ such that for all j and $i = 1,2$, $T_i^{(j+1)}$ is a Morita extension of $T_i^{(j)}$ and $T_1^{(n)}$ is logically equivalent to $T_2^{(m)}$. Informally, and slightly inaccurately, we will say that T_1 and T_2 are Morita equivalent just in case they have a common Morita extension, here given by $T_1^{(n)}$.

With the technical details in place, let us turn towards examining Barrett and Halvorson's argument. In outline, their argument can be construed along the following lines:

(1) If T is a many-sorted theory and T^+ is a Morita extension of T, then T and T^+ have the same ontology
(2) Any geometry meeting certain natural conditions can be axiomatized Morita equivalently either in terms of points, as a theory T_P, or in terms of lines, as a theory T_L
(3) Therefore, the alternative axiomatisations of any such geometry have the same ontology
(4) Therefore, since conceptual relativity requires that the theories involved have incompatible ontologies, geometries that treat lines as basic and geometries that treat points as basic are not an instance of conceptual relativity.

There are two immediate worries about this argument. The first is that (1) seems implausible: intuitively, Morita extensions have an expanded ontology relative to their base theories because they introduce new sorts. We will turn to this concern in due course. The second worry is that Barrett and Halvorson swindle at the outset by axiomatizing the point-based and line-based geometries in ways *that fail to be incompatible*. Recall that conceptual relativity involves theories that are jointly inconsistent, at least as long as it is assumed that the terms that occur in both theories are to receive one and the same interpretation across the two theories. That is precisely what we saw in the axiomatizations ZFCP and ZFCL in Section 3.1: taken together, these theories are

syntactically inconsistent because ZFCP defines lines as sets of points and ZFCL defines points as sets of lines, which in conjunction contradicts the Axiom of Foundation. The axiomatizations T_P and T_L are jointly consistent however, as can be seen from the fact that they have a consistent common Morita extension T^+ such that $T^+ \vdash T_P \wedge T_L$. This raises the worry that Barrett and Halvorson are basing their argument on axiomatizations that fail to capture the incompatibility crucial to the example's being an instance of conceptual relativity. One might allege that Barrett and Halvorson have trivialised the compatibility of the two theories for which they are trying to argue.

In defence of Barrett and Halvorson, note however that the theories T_P and T_L they have chosen are formalizations of perfectly standard textbook axiom systems for geometry. If standard axiomatizations of geometry in terms of points and lines turn out to be jointly consistent, then that is a problem for Putnam's argument, not for Barrett and Halvorson's. Putnam alleges that our theorizing often yields "equivalent but incompatible" descriptions of the world. For this to be of interest, it ought to be evidenced by theories that occur in our ordinary mathematical and scientific practice. Therefore, if standard axiomatizations of geometry in terms of points and lines turn out to be compatible, that emphatically does not mean Barrett and Halvorson have to shoehorn them into being jointly inconsistent so that they fit Putnam's criteria for conceptual relativity. Rather, it should make us alert to the fact that the air of inconsistency we get from equivalent formalisations of the same theory that differ by which terms they treat as basic and which as defined is often misleading. It arises due to a mistaken assumption that theories claim of the terms they treat as primitive that they denote fundamental entities. Thus, T_P and T_L might strike us as inconsistent because we assume that T_P asserts that points are fundamental and lines derivative whereas T_L asserts the opposite. In fact however, neither theory asserts any such thing. T_P is consistent with lines being fundamental just as T_L is consistent with points being fundamental.[105] That is why even if metaphysicians established to everyone's satisfaction that points, say, are fundamental, T_L would remain a perfectly correct theory. Claims of relative fundamentality are usually not built into our mathematical and scientific theories, precisely because they tend to raise otiose and irrelevant issues. What is really of interest then is not whether T_P and T_L are compatible but whether they are fully interchangeable. Insofar as T_P and T_L are

[105] A linguistic piece of evidence for this comes in the form of a cancellability test: 'I accept T_L but do not believe that lines are fundamental' is a perfectly consistent thing to say and does not sound at all odd, which shows that any implication carried by acceptance of T_L that lines are fundamental is separable from commitment to T_L and thus cancellable.

natural and seemingly equally suitable axiomatizations of geometry, it remains an important and non-trivial task to show that they are equivalent and that therefore the choice between them is not a substantive one that would involve having to choose between genuinely distinct ontologies. It is this that Barrett and Halvorson's argument is supposed to establish.

Crucial to their argument is claim (1), that Morita extensions have the same ontology as their underlying base theories. On the face of it, this seems like it couldn't possibly be true for the simple reason that constructing new sorts introduces new domains of quantification. To put it crudely, if T^+ is a Morita extension of T, each model of T^+ will contain more stuff than the corresponding model of T. So how could the two theories possibly have the same ontology? This leads to a further worry: if the content of a Morita extension T^+ goes beyond that of the base theory T, then two theories' T_1 and T_2 having a common Morita extension does not guarantee their equivalence, since it might be that T_1 and T_2 say different things while T^+ entails the conjunction of what T_1 and T_2 say. Unlike a common core theory then, a common Morita equivalence would not warrant a judgment of equivalence. In defence of (1), Barrett and Halvorson offer what I will call the *paraphrase argument*. I will argue that while this argument faces certain problems, it can be made to work. Barrett and Halvorson's claim that Morita equivalent theories have the same ontology is therefore in good standing.

3.2.1 The paraphrase argument

In defence of the claim that a Morita extension has the same ontology as the theory it extends, and indeed that the two are equivalent, Barrett and Halvorson [2017] appeal to the following result:

Proposition 1. Let T^+ be a Morita extension of T and suppose $\sigma \in \Sigma(T^+)$ is constructed in terms of sorts $\sigma_1, \sigma_2 \in \Sigma(T)$. Then, for any formula $\psi(x) \in \mathcal{L}(T^+)$ with x of sort σ, there exists a formula $\phi(y,z) \in \mathcal{L}(T)$ with y, z of sorts σ_1 and σ_2, respectively, such that

$$T^+ \vDash \exists_\sigma x \psi(x) \leftrightarrow \exists_{\sigma_1} y \exists \sigma_2 z \phi(y,z) \tag{7.6}$$

In a nutshell, Proposition 1 states that every existential sentence in the language of a Morita extension can be paraphrased as an existential sentence that only mentions the sorts of the base theory. Proposition 1 therefore captures a sense in which any existence claim made in the extended language depends for its truth wholly on the existence and properties of the objects in the original theory's intended domain of quantification. Saying that however still does not give us a full argument for the claim that Morita extensions have the same ontology as their underlying base theory. It is moreover not straightforward to extract a clear argument from what Barrett and Halvorson say. Broadly, there are two ways to construct an argument to this effect: first, one can argue that Morita extensions do not have an increased ontological commitment relative to their base theories because they are not really committed to the new sorts they appear to be quantifying over. Call this the *minimalist's argument*. Secondly, one can argue that Morita extensions do not have an increased ontological commitment because the base theory over which they are constructed is implicitly *already* committed to the sorts of objects the Morita extension introduces. Call this the *maximalist's argument*. In their discussion, Barrett and Halvorson vacillate between these two options so that it never becomes entirely clear which of them they intend to endorse. However, for several reasons the minimalist's argument is the more natural one to rest on Proposition 1. We will therefore consider it first.

The minimalist's argument is a natural one to develop on the basis of Proposition 1. After all, the intuitive import of this result is that any quantification over objects of the newly introduced sorts is really just quantification over objects of the old sorts in disguise. Therefore, the argument will go, ontological commitment to objects of the new sorts is merely apparent. In a slogan, commitment to objects of the new sorts can always be "paraphrased away". Barrett and Halvorson seem to have this minimalist construal of the argument in mind when they state that when we construct a subsort of a sort σ, "the new sort [...] does not contain ''new objects'' that are independent of the old objects. It instead just provides us with a new way of talking about some of the objects of sort σ." (Barrett and Halvorson [2017], p. 1057) And of course, paraphrase is a well-known standard tool one will employ in arguments to the effect that a certain ontological commitment is merely apparent. To see this style of argument in action, it will be helpful to look at a simple example. Consider the following argument (see Keller [2017]):

(1) There is a crack in the vase
(2) Therefore, there are cracks

The argument appears valid and has a seemingly innocuous premiss, but yields a potentially troublesome conclusion. After all, (2) is naturally read as asserting the existence of cracks. But cracks are perhaps not the sort of thing we are happy to include in our ontology. Paraphrase promises a way out of this trouble. One could argue for instance that (1) says no more than and hence can be paraphrased as (1*) 'The vase is cracked'. Substituting (1*) for (1) yields

(1*) The vase is cracked
(2) Therefore, there are cracks

This argument, unlike the first, does not feel intuitively valid. So, the suggestion goes, we can think of (1) as a potentially misleading way of expressing (1*). Once it has been clarified through paraphrase that this is so, the validity of the first argument is revealed to have been an illusion (Keller [2017]).

In fact, this is not the only way an appeal to paraphrase may be thought to help with avoiding ontological commitment to cracks. Alternatively, or additionally, one might argue that (2) can be paraphrased as (2*) 'Some things are cracked'. This yields the following reconstruction of the argument:

(1*) The vase is cracked
(2*) Therefore, some things are cracked

That version of the argument looks valid, but no longer appears to have any troublesome ontological implications. In short, the role of paraphrase in such arguments is to show that we do not have to take ourselves to be committed to certain dubious entities because any assertion that appears to be quantifying over them can be paraphrased in a way that eliminates that quantification.

Paraphrase arguments of this kind operate against a broadly Quinean account of ontological commitment. According to this account, one's ontological commitments are not

supposed to be overly difficult to determine: they can be read off the existence claims one either accepts or can deduce from sentences one accepts (Quine [1948]). For instance, I am ontologically committed to dogs because I accept the truth of 'There are dogs'. But the story doesn't end there. Some of our (apparent) ontological commitments are unwanted. To give a well-worn example, a nominalist might not want to accept that numbers exist, even though he believes that there exists a prime number between 10 and 100; a belief that seems to entail that numbers exist. As a way around undesired ontological commitments, Quine recommends precisely the tool of paraphrase we have been considering. Thus, he says:

> [W]hen we say that some zoological species are cross-fertile we are committing ourselves to recognizing as entities the several species themselves, abstract though they are. We remain so committed at least until we devise some way of so paraphrasing the statement as to show that the seeming reference to species on the part of our bound variable was an avoidable manner of speaking.
>
> (Quine [1948], p. 39)

But once we appreciate the Quinean background against which paraphrase arguments operate, certain problems for the minimalist construal of Barrett and Halvorson's argument become apparent. To appreciate the first such problem, it will help to recall another Quinean tenet, namely that our theories ought to be regimented in a formal language (specifically, a first order language, of course). Quine [1954b] recommends that we translate our theories into a formal language for the purpose of enhancing clarity, dissolving ambiguities, and fixing entailment relations; in other words, to rectify problems that are pervasive in ordinary language. Indeed, one way to think about the crack argument from above is precisely as an instance of confusion with regard to our ontological commitments that arises from ambiguity in and unclear entailments between sentences in ordinary language. The paraphrases we have offered in effect amount to different suggestions for how to regiment the argument in a formal language. The insight to be gained from proceeding in this way is that for every regimentation of the argument, either the argument is construed as invalid, or the first premiss is rendered implausible, or the conclusion is shown to be ontologically innocuous. Regimentation then is really a thorough kind of paraphrase that systematically roots out ambiguity and unclarity.

Conversely, this means of course that theories that are already regimented are not the kind of theories one should take to exhibit ambiguity or unclear entailment relations. Applied to the question of a theory's ontological commitments once more, this means that we should take a regimented theory T to be ontologically committed to a kind of thing ψ just in case

$$T \vdash \exists x \psi(x) \tag{7.7}$$

One problem with the minimalist's paraphrase argument therefore is that it is meant to apply to sentences of a formal language even though such sentences are supposed to wear their ontological commitments "on their sleeves" already, as it were.

There arises also a second problem. In the minimalist's hands, Proposition 1 is employed in argument as follows:

(1) If T^+ is a Morita extension of T then by Proposition 1, every existential sentence in the language of the Morita extension is equivalent to an existential sentence in the language $\mathcal{L}(T)$
(2) Therefore, T^+ is ontologically committed at most to the objects in T's intended domain of quantification

But this argument leads to apparent absurdity: Let T^+ be the common Morita extension of our Morita equivalent axiomatisations of geometry T_P and T_L. By (1), for every existential sentence S of $\mathcal{L}(T^+)$ there exist sentences S_P from $\mathcal{L}(T_P)$ and S_L from $\mathcal{L}(T_L)$ such that S is equivalent modulo T^+ to S_P and S_L, respectively. By (2), T^+ is not ontologically committed to points since T_L's domain of quantification does not include points. Furthermore, T^+ is not ontologically committed to lines, since T_P's domain of quantification does not include lines. So, T^+ is not ontologically committed to either points or lines. This already seems absurd. Moreover, given that S is equivalent to each of its paraphrases S_P and S_L, it follows that S_P and S_L are equivalent. Since S is arbitrary, it follows that every existential sentence from $\mathcal{L}(T_P)$ can be paraphrased in $\mathcal{L}(T_L)$, and vice versa. If we follow the minimalist's argument, we must conclude that T_L isn't committed to lines and T_P isn't committed to points. That seems even more absurd!

All in all then, the situation looks rather dire for the minimalist: first, we noted that his style of "eliminative" paraphrase argument usually applies to theories formulated in natural, unregimented language only. Secondly, we saw that his reasoning leads to apparent absurdity. There are two ways forward: we can try to salvage the minimalist's argument or we can opt for the maximalist's argument. The second option amounts to arguing that theories by default include in their ontology all objects constructible from the objects they treat as primitive. Since points are constructible from lines and lines from points, each of T_P and T_L is committed to both points and lines, the argument goes. Clearly, these two

approaches pull in opposite directions. Still, I think that much can be said for either option. While I am ultimately more sympathetic to the maximalist view, the minimalist view turns out to be more resilient than one might think in light of the problems we identified.

3.2.2 The Paraphrase argument saved

Let us begin with the minimalist approach. The idea here is to salvage the paraphrase argument. There were two concerns. The first was that paraphrase arguments should only be applied to unregimented theories since regimented theories already perspicuously display their ontological commitments. To counter this concern, Barrett and Halvorson can point to the specific function Morita extensions serve: one way to understand them is as giving us convenient constructions for talking about the objects of the original theory's sorts in a more compact, less cumbersome manner. But when they are constructed for this purpose, the resulting theory is optimized for expressive economy, not ontological clarity. In other words, Morita extensions differ from regimentations of natural language theories in their aims. There can be no presumption then that one can read the ontological commitments of a Morita extension directly off the existential sentences it proves. This means that there remains a place for paraphrase arguments.

The second, more serious worry however was that the paraphrase argument shaves off too much ontological commitment: we reached the conclusion that the common Morita extension T^+ of T_P and T_L was committed to neither lines nor points; indeed that none of the three theories was committed to either lines or points. And that seemed absurd. In fact however, it is not absurd at all. To see why this is so, we will have to get somewhat clearer on the notion of ontological commitment. Wilson [1981] helpfully points out that Quine draws a distinction between a theory's ontology and its ontological commitments. A theory's ontology is the domain of objects it describes, whether correctly or incorrectly. Thus, the ontology of Peano Arithmetic is the natural numbers for instance. The ontology of ZFC is the set-theoretic universe, etc. A theory's ontological commitment however can come apart from its ontology. Quine defines ontological commitment as follows:

> A theory is committed to those and only those entities to which the bound variables of the theory must be capable of referring in order that the affirmations made in the theory be true.
>
> (Quine [1948], p. 33)

As Wilson explains, two theories can have the same ontology while differing in their ontological commitments. ZFC and ZFC+¬CH each have as their ontology the set-theoretic universe. The latter however is committed to the existence of certain sets to which the former is not committed; specifically, sets of a cardinality strictly between that of the natural numbers and that of the power set of the natural numbers.

Now, Quine's definition of ontological commitment as applied to the case of T_P, T_L, and their common Morita extension T^+ tells us that each theory is committed only to those entities that must exist for the theory to be true. But then it simply turns out that none of the three theories is committed to points, nor to lines – the conclusion we initially deemed absurd. The reason for this is the following: suppose the mathematical realm is such that there are only points – perhaps God ordained it so. Lines are a mere figment of human imagination; something we construct in our minds because it simplifies our theorizing. Due to the intertranslatability of T_P and T_L, T_L is nonetheless able to provide a perfectly adequate description of this realm. Every sentence in T_L can be interpreted, thanks to the translation function, as a paraphrase of some claim about points. But then, T_L can be perfectly true even if lines don't exist. In other words, T_L is not ontologically committed to lines. By parity of reasoning, T_P is not committed to points, and neither is the common Morita extension T^+. The conclusion we reached via the minimalist's paraphrase argument is perfectly correct, even though it may have seemed incredible at the start. Of course, at least one of the two sorts of entities, points or lines, will have to exist in order for the three theories to be true. In other words, the three theories are committed to there being *either points or lines*. But that is all, and it is consistent with the conclusion of the minimalist's argument.

So, Morita equivalence provides a philosophically satisfactory standard of equivalence that motivationalists can adopt. The intertranslatability of T_P and T_L that we get from the Morita construction shows the two theories to be related by a horizontal reformulation: neither one contains any surplus structure relative to the other. The motivationalist can therefore regard the two theories as equivalent without having to find a further common core theory – the theory already is all core, as it were. Of course, as we have noted repeatedly, motivationalists can maintain that one of the formulations T_P and T_L is more ontologically perspicuous or joint-carving than the other. But they need not do so. They can go also go along with Putnam and Wallace in thinking that it makes no sense to ask which of the two is more perspicuous.

3.2.3 Maximalism

While the minimalist's paraphrase argument is internally consistent, as I have attempted to show, one may nonetheless feel a dissatisfaction with the minimalist's view. According to the minimalist, only what is fundamental is real. The ontology of our geometries is therefore either just points or just lines. This view seems to go against the prevailing sentiment that mathematical ontology is vast and plenitudinous: at the very least, whatever is constructible from mathematical objects is a further mathematical object. Call this the maximalist's view. The maximalist, too, argues that T_P and T_L have the same ontology. But he thinks that they are each committed to both points and lines. This maximalist sentiment finds expression at certain points in Barrett and Halvorson's discussion. Thus, they state that

> the common Morita extension T^+ is a theory that quantifies over both points and lines. The theories T_P and T_L are simply convenient ways of expressing the geometric facts that are more fully expressed by the comprehensive theory T^+.
>
> (Barrett and Halvorson [2017], p. 1060)

Maximalists will think about the paraphrase argument based on Proposition 1 differently from minimalists. Proposition 1 captures a sense in which the point sort and the line sort are each eliminable in the presence of the other: all that can be said in the language of lines can be said in the language of points, and vice versa. The minimalist thinks about this eliminability ontologically: only one of the sorts is real, at least fundamentally. The maximalist however attributes importance to this eliminability result because it shows that everything that is said by T^+ is also said (albeit perhaps in different words) by each of T_P and T_L. This means that T_P and T_L say exactly the same things – they are equivalent. For the maximalist, the significance of Proposition 1 is therefore that it shows common Morita extensions to be theories that are apt to warrant judgments of equivalence. We know that T_P and T_L are equivalent because T^+ says precisely what T_P says and precisely what T_L says.

For motivationalists, the ability to warrant judgments of equivalence for theories is a necessary feature of common core theories. We see therefore that Morita extensions meet this important criterion for common core theories. And indeed, on the maximalist view, the common Morita extension will be the preferred formulation out of the three, since it explicitly displays the ontological commitments shared between the three formulations.

Combining motivationalism with maximalism raises one final issue however that needs to be addressed. Since the process of sort construction and Morita extension can be iterated indefinitely, even the common Morita extension of T_P and T_L does not fully display the objects it is committed to. This raises the question of whether we can ever give a theory that comprehensively delineates the ontology the maximalist takes himself to be committed to, i.e. whether we can construct a fully ontologically perspicuous theory of the kind motivationalists view as ideal. In fact however, this isn't very hard to do. To show how it can be done, we will need to introduce the notion of a maximal Morita extension. Intuitively, a maximal Morita extension T* of a theory T is a Morita extension that defines all of the sorts constructible from the sorts in T. More formally:

Definition (Maximal Morita Extension): Let T be a many-sorted theory. Let T* be a Morita extension of T and suppose that for all sorts $\sigma, \sigma' \in \mathcal{L}(T)$ such that σ, σ' admit the construction of a new sort, there exists $\sigma^* \in \mathcal{L}(T^*)$ with the corresponding definitional formula $\delta_{\sigma^*}(\sigma, \sigma') \in T^*$. Then T* is a *maximal Morita extension* T* of T.

Evidently, for any given T, T^* is unique up to definitional equivalence. In other words, if T^* and $T^{*\prime}$ are two maximal Morita extensions of a theory T, then they are definitionally equivalent. With the notion of a maximal Morita extension in hand, we can define what we will call a Super-Morita extension T^S of a theory T.

$$T^S = T \cup \bigcup_{n \in \mathbb{N}} T_n^* \tag{7.8}$$

Where T_1^* is a maximal Morita extension of T and for all n, T_{n+1}^* is a maximal Morita extension of T_n. We may then take T^S to capture the maximalist's ontological commitments. One might wonder of course about Morita extensions of T^S. Would we not have to build those as well if we want to go for a truly maximal ontology? And if so, does the worry about giving a comprehensive theory of the maximalist's ontological commitments not reappear? This worry is unfounded. T^S is a fixed point of the Morita construction in a precise sense:

Proposition 2. Let T^S be a Super-Morita extension of a theory T and let T^{S+} be a Morita extension of T^S. Then T^S and T^{S+} are definitionally equivalent.

Proof: Let σ be an arbitrary sort in $\Sigma(T^{S+}) \backslash \Sigma(T^S)$, where for a theory T, $\Sigma(T)$ denotes the signature of T. Then T^{S+} defines σ in terms of sorts τ, ρ $\in \Sigma(T^S)$. By the definition of T^S,

there is some $n \in \mathbb{N}$ such that $\tau, \rho \in \Sigma(T_n^*)$. But then T_{n+1}^* defines σ', where σ' and σ are interdefinable. Q.E.D.

4 Conclusion

In this chapter, I have attempted to show that the common core account of equivalence provides motivationalists with a powerful tool for dealing with alleged cases of strong conceptual relativity. The upshot is that the phenomenon of conceptual relativity should seem much less threatening to motivationalism than perhaps it did initially. If examples of plausibly equivalent theories that cannot be reconciled by common core methods were pervasive, motivationalism would seem too demanding a view. As it stands however, this worry lacks a solid foundation. Of course, there might be genuine cases of equivalent theories that do not have a common core. But even that is not necessarily a problem for motivationalism. Plausibly, motivationalists can maintain that without a common core, we will always lack the appropriate warrant for interpreting such theories as equivalent. They may be equivalent, but we will never know for certain.

8 Conclusion

My aim in this thesis has been to develop and defend a form of motivationalism about symmetries and dualities. I hope to have articulated a principled motivationalism with clear, well-defined criteria for when it is justified to interpret symmetry-related or dual models as equivalent, but one that is not implausibly demanding. Like all motivationalists, I maintain that one is not licensed to interpret symmetry-related or dual models as equivalent simply because they are empirically equivalent. Like most motivationalists, I believe that reformulation plays a crucial role in our interpretation of such models. Given the importance I assign to reformulation in theory interpretation, the question arises how best to go about reformulating theories. I have aimed to show that reduction is a more fruitful and versatile approach to reformulation than is often thought. In consequence, sophistication cannot be assumed to be superior across the board. Both strategies should be pursued and comparisons will have to be made case by case.

The views defended in this thesis were shaped by the pressure of anti-motivationalist argument. What emerges is a position more moderate than standard motivationalism, but in my estimation also more plausible and more precise. Indeed, I take the main insight of this thesis to be that a squarely motivationalist position survives once all the most important criticisms have been factored in. Many motivationalists maintain that before we are permitted to interpret any two empirically equivalent models as also physically equivalent, we need a clear, coherent account of the one description they supposedly give of the world; a "coherent metaphysical picture of the common ontology underpinning their equivalence" (Martens and Read [2020], p. 320). Some, like Jacobs, even insist that this picture must be provided in the form of a fully intrinsic common core model. The fact that judgments of equivalence often rightly precede a full interpretation of the models in question and a full articulation of this metaphysical picture is commonly thought to be a serious problem for this view.

As a first step towards a solution, I proposed that we rethink the putatively very tight connection between constructing a common core model and providing the sought metaphysical picture. Not every common core tells a metaphysical story, and not every specification of a shared ontology amounts to a full specification of a mathematical model. Since the two can come apart, their importance for judgments of equivalence may also

differ. I have argued that it is the common core, but not the metaphysical picture, that must be given to warrant a judgment of equivalence. Divorcing the two opens the way for a salutary moderate motivationalism along two dimensions: substantial and formal. On the side of substance, we note that the common core need not be intrinsic or fully perspicuous. It can retain the interpretive difficulties of the models of which it is a part and even create new ones (although it needs to be assessable whether the core model has explanatory deficiencies as compared to the original models). Moreover, it can carry surplus structure of its own and thus need not be "the final word", as it were – a perfectly artifact-free representation of the shared fundamental ontology. Concomitantly, on the formal side, the common core can be constructed extrinsically, using auxiliary structure. While a perfectly intrinsic characterization of the common core in the style of Hilbert or Tarski may be an ideal, it is not a requirement. This gives us a motivationalism that is attuned to standards of equivalence accepted by working mathematicians and physicists. Motivationalists need not be purists or fiery-eyed ascetic preachers who for the sake of intrinsicality demand a full first-order axiomatization of all of our theories (they can be of course, but I would advise against it)

Hand in hand with this goes a weakening of the motivationalist standard of formal equivalence for models. While many motivationalists regard isomorphism as the only mathematical relation between models that immediately underwrites judgments of equivalence and beyond which no further justification for equivalence has to be offered, I have argued that we should adopt mathematical equivalence as our standard of equivalence. This can include standards such as Morita equivalence and more informal standards of 1:1 intertransformability of the kind mentioned by Wallace [2022a]. There is a worry that mathematically equivalent formulations can suggest different metaphysical pictures, but I have argued that judgments of equivalence can be warranted even if we have not determined one metaphysical picture as the settled interpretation. This is not to say that there are no important interpretive questions regarding mathematically equivalent formulations of course. None of what I have argued is meant to suggest that we cannot or should not ask which representation of a theory is most perspicuous, which structure is directly representational, which structure merely encodes information in a more indirect way, and so forth.

Even with weak demands on common cores and a wider standard of formal equivalence, there remains a worry that motivationalism is overly demanding. Those who believe in conceptual relativity maintain that fact and convention will inevitably be intertwined in our theories to such an extent that we cannot hope to always find a common core free of representational artifacts for all pairs of equivalent theories. My response to this is two-pronged. First, the common core approach is more powerful than is often thought and can account for theoretical equivalence in surprisingly many cases. But secondly, even if the defender of conceptual relativity is right that certain equivalent theories do not have a common core, this does not in itself undermine motivationalism. Motivationalism is a view on the epistemology of equivalence. It tells us when our beliefs that certain theories are equivalent are warranted. What is required is that we produce a common core theory. So, even if in fact there are equivalent theories without a common core, motivationalists can maintain that in these cases, we can never be certain that the theories really are equivalent. The reason for this is that we cannot tell whether all the structural differences between the theories are entirely due to artifacts of representation. As we saw, in many cases judgments of equivalence believers in conceptual relativity are prepared to make turn out to have been premature. This lends support to the more cautious view of the motivationalist.

Overall, this thesis has aimed to formulate and defend a unified motivationalism that applies the same standards of judgment to both intra- and intertheoretic equivalences, i.e., symmetries and dualities. Care was taken to incorporate insights of interpretationalists and advocates of formal approaches to equivalence. The view that emerges however remains a principled, well-delineated form of motivationalism.

References

Albert, David (2013). Wave function realism. In A. Ney and D. Albert (Eds), The Wave Function: Essays on the Metaphysics of Quantum Mechanics, pp. 52–7. Oxford: Oxford University Press

Alexander, H. G. (ed.) (1956). The Leibniz-Clarke Correspondence. Manchester University Press.

Andréka, Hajnal ; Madarász, Judit ; Németi, István & Székely, Gergely (2024). Testing Definitional Equivalence of Theories Via Automorphism Groups. Review of Symbolic Logic 17 (4):1097-1118.

Arntzenius, Frank (1990). Causal paradoxes in special relativity. British Journal for the Philosophy of Science 41 (2):223-243.

Arntzenius, Frank (2012). Space, time, & stuff. New York: Oxford Univ. Press. Edited by Cian Seán Dorr.

Baker, David John (2020). Some Consequences of Physics for the Comparative Metaphysics of Quantity. In Karen Bennett & Dean W. Zimmerman (eds.), Oxford Studies in Metaphysics Volume 12. Oxford University Press. pp. 75-112.

Baker, David John (unpublished). Comparativism with mixed relations. philsci-archive.pitt.edu/20814/

Barbour, Julian (1999). The End of Time: The Next Revolution in Physics. Weidenfeld & Nicholson.

Barbour, Julian (2001). On general covariance and best matching. in Callender, Craig & Huggett, Nick (2001). Physics Meets Philosophy at the Planck Scale: Contemporary Theories in Quantum Gravity. Cambridge University Press.

Barbour, Julian B. (2003). Scale-invariant gravity: Particle dynamics. Classical and Quantum Gravity 20:1543--70.

Barbour, Julian B. & Bertotti, Bruno (1982). Mach's principle and the structure of dynamical theories. Proceedings of the Royal Society, London:295--306.

Baron, Sam; Le Bihan, Baptiste & Read, James (2025). Scientific Theory and Possibility. Erkenntnis 1:1-17.

Barrett, Thomas William (2022). How to count structure. Noûs 56 (2):295-322.

Barrett, Thomas William & Halvorson, Hans (2016a). Glymour and Quine on Theoretical Equivalence. Journal of Philosophical Logic 45 (5):467-483.

Barrett, Thomas William & Halvorson, Hans (2016b). Morita Equivalence. Review of Symbolic Logic 9 (3):556-582.

Barrett, Thomas William & Halvorson, Hans (2017). From Geometry to Conceptual Relativity. Erkenntnis 82 (5):1043-1063.

Barrett, Thomas William & Halvorson, Hans (2022). Mutual translatability, equivalence, and the structure of theories. Synthese 200 (3):1-36.

Barrett, Thomas William ; Manchak, J. B. & Weatherall, James Owen (2023). On automorphism criteria for comparing amounts of mathematical structure. Synthese 201 (6):1-14.

Becker, K., Becker, M., & Schwarz, J. (2007). String theory and M-theory: A modern introduction. Cambridge: Cambridge University Press

Belot, Gordon (1998). Understanding electromagnetism. British Journal for the Philosophy of Science 49 (4):531-555.

Belot, Gordon (1999). Rehabilitating relationalism. International Studies in the Philosophy of Science 13 (1):35 – 52.

Belot, Gordon (2000). Geometry and motion. British Journal for the Philosophy of Science 51 (4):561--95.

Belot, Gordon (2002). Symmetry and gauge freedom. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 34 (2):189-225.

Belot, Gordon (2013). Symmetry and Equivalence. In Robert Batterman, The Oxford Handbook of Philosophy of Physics. Oxford University Press USA. pp. 318-339.

Belot, Gordon (2018). Fifty Million Elvis Fans Can't be Wrong. Noûs:946-981.

Benitez, Federico (2019). Selective Realism and the Framework/Interaction Distinction: A Taxonomy of Fundamental Physical Theories. Foundations of Physics 49 (7):700-716.

Bigelow, John ; Pargetter, Robert & Armstrong, D. M. (1988). Quantities. Philosophical Studies 54 (3):287 - 304.

Bokulich, Alisa (2008). Can classical structures explain quantum phenomena? British Journal for the Philosophy of Science 59 (2):217-235.

Bokulich, Alisa (2011). How scientific models can explain. Synthese 180 (1):33 - 45.

Bokulich, Alisa (2016). Fiction As a Vehicle for Truth: Moving Beyond the Ontic Conception. The Monist 99 (3):260-279.

Bokulich, Alisa (2020). Losing Sight of the Forest for the Ψ: Beyond the Wavefunction Hegemony. In Juha Saatsi & Steven French, Scientific Realism and the Quantum. Oxford: Oxford University Press.

Brading, Katherine & Castellani, Elena (eds.) (2002). Symmetries in Physics: Philosophical Reflections. New York: Cambridge University Press.

Bradley, Clara (2021). The Non-equivalence of Einstein and Lorentz. British Journal for the Philosophy of Science 72 (4):1039-1059.

Brown, Harvey (2005). Physical Relativity: Space-Time Structure From a Dynamical Perspective. Oxford, GB: Oxford University Press UK.

Brown, Harvey R. & Read, James (2022). The Dynamical Approach to Spacetime Theories. In Eleanor Knox & Alastair Wilson, The Routledge Companion to Philosophy of Physics. London, UK: Routledge.

Burgess, John P. (2005). Being Explained Away. The Harvard Review of Philosophy 13 (2):41-56.

Burgess, John P. & Rosen, Gideon (1997). A subject with no object: strategies for nominalistic interpretation of mathematics. New York: Oxford University Press. Edited by Gideon A. Rosen.

Burgess, John P. & Rosen, Gideon (2005). Nominalism Reconsidered. In Stewart Shapiro, Oxford Handbook of Philosophy of Mathematics and Logic. Oxford and New York: Oxford University Press.

Butterfield, Jeremy (1989). The hole truth. British Journal for the Philosophy of Science 40 (1):1-28.

Butterfield, J. (2021). On dualities and equivalences between physical theories. In C. Wüthrich, B. Le Bihan, and N. Huggett (Eds.), Philosophy Beyond Spacetime: Implications from Quantum Gravity, pp. 41–77. Oxford: Oxford University Press.

Callender, Craig & Huggett, Nick (2001). Physics Meets Philosophy at the Planck Scale: Contemporary Theories in Quantum Gravity. Cambridge University Press.

Capozziello, Salvatore, Vittorio De Falco, and Carmen Ferrara (2022). "Comparing equivalent gravities: common features and differences." Eur. Phys. J. C 82 (10): 865

Case, Jennifer (2001). The heart of Putnam's pluralistic realism. Revue Internationale de Philosophie 4:417-430.

Caulton, Adam (2015). The role of symmetry in the interpretation of physical theories. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 52 (Part B):153-162.

Chen, Elliott D. (2023). Newtonian gravitation in Maxwell spacetime. Studies in History and Philosophy of Science Part A 102 (C):22-30.

Chomsky, Noam (1957). Syntactic Structures. Mouton.

Coffey, Kevin (2014). Theoretical Equivalence as Interpretative Equivalence. British Journal for the Philosophy of Science 65 (4):821-844.

Coffey, Kevin (2024). (Competing?) Formulations of Newtonian Gravitation. Journal of Philosophy 121 (11):628-656.

Curiel, Erik (2018). On the Existence of Spacetime Structure. British Journal for the Philosophy of Science 69 (2):447-483.

Dasgupta, Shamik (2011). The bare necessities. Philosophical Perspectives 25 (1):115-160.

Dasgupta, Shamik (2013). Absolutism vs Comparativism About Quantity. Oxford Studies in Metaphysics 8:105-150.

Dasgupta, Shamik (2016). Symmetry as an Epistemic Notion. British Journal for the Philosophy of Science 67 (3):837-878.

De Haro, S. (2016). Spacetime and Physical Equivalence. in Space and Time after Quantum Gravity, Huggett, N. and Wüthrich, C. (Eds.), http://philsci-archive.pitt.edu/13243/

De Haro, Sebastian (2019a). The heuristic function of duality. Synthese 196 (12):5169-5203.

De Haro, Sebastian (2019b). Theoretical equivalence and duality. Synthese 198 (6):5139-5177.

De Haro, Sebastian (2021). The Empirical Under-Determination Argument Against Scientific Realism for Dual Theories. Erkenntnis 88 (1):117-145.

De Haro, Sebastian & Butterfield, Jeremy (2021). On symmetry and duality. Synthese 198 (4):2973-3013.

De Haro, Sebastian & Butterfield, Jeremy (2025). The Philosophy and Physics of Duality. OUP

De Haro, Sebastian ; Teh, Nicholas & Butterfield, Jeremy N. (2017). Comparing dualities and gauge symmetries. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 59:68-80.

Dewar, Neil (2018). Maxwell Gravitation. Philosophy of Science 85 (2):249-270.

Dewar, Neil (2019). Sophistication about Symmetries. British Journal for the Philosophy of Science 70 (2):485-521.

Dewar, Neil (2022). Structure and Equivalence. Cambridge University Press.

Dewar, Neil (2023a). Interpretation and equivalence; or, equivalence and interpretation. Synthese 201 (4):1-24.

Dewar, Neil (2023b). On Internal Structure, Categorical Structure, and Representation. Philosophy of Science 90 (1):188-195.

Dewar, Neil (2024). On Absolute Units. British Journal for the Philosophy of Science 75 (1):1-30.

Dewar, Neil and Eva, Benjamin (unpublished). A Categorical Perspective on Symmetries and Equivalence

Dougherty, John (2021). I ain't afraid of no ghost. Studies in History and Philosophy of Science Part A 88 (C):70-84.

Dougherty, John (forthcoming). Effective and Selective Realisms. British Journal for the Philosophy of Science.

Dirac, Paul (1930). The principles of quantum mechanics. Oxford,: Clarendon Press.

Dürr, D., S. Goldstein and N. Zanghi, "Bohmian Mechanics as the Foundations of Quantum Mechanics", in J. T. Cushing et al., (eds.), Bohmian Mechanics and Quantum Theory: An Appraisal, Kluwer Academic Publishers, 1996.

Dürr, Detlef, Sheldon Goldstein, and Nino Zanghì (1997), "Bohmian Mechanics and the Meaning of the Wave Function", in R.S. Cohen, M. Horne, and J. Stachel (eds), *Experimental Metaphysics—Quantum Mechanical Studies for Abner Shimony, Volume One*, (Boston Studies in the Philosophy of Science 193), Boston: Kluwer Academic Publishers.

Dürr, Patrick and Read, James (2023). Reconsidering Conventionalism: An invitation to a sophisticated philosophy for modern (space-)times [Preprint]

Earman, John (1989). World Enough and Spacetime. MIT press.

Earman, J. (1995). Bangs, Crunches, Whimpers, and Shrieks: Singularities and Acausalities in Relativistic Spacetimes. New York: Oxford University Press.

Fara, Delia Graff (2009). Dear haecceitism. Erkenntnis 70 (3):285–297.

Field, Hartry H. (1980). Science Without Numbers: A Defence of Nominalism. Princeton, NJ, USA: Princeton University Press.

Fletcher, Samuel C. ; Manchak, J. B. ; Schneider, Mike D. & Weatherall, James Owen (2018). Would two dimensions be world enough for spacetime? Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 63:100-113.

Friedman, Michael (2001). Dynamics of reason: the 1999 Kant lectures at Stanford University. Stanford, Calif.: CSLI Publications.

Ghirardi, G., A. Rimini and T. Weber (1986). Unified Dynamics for Micro and Macro Systems. Physical Review D 34: 470-491

Glymour, Clark (1977). The epistemology of geometry. Noûs 11 (3):227-251.

Grimmer, Daniel, Cinti, Enrico & Jaksland, Rasmus (forthcoming). Duality, Underdetermination, and the Uncommon Common Core. British Journal for the Philosophy of Science.

Halvorson, Hans (2012). What Scientific Theories Could Not Be. Philosophy of Science 79 (2):183-206.

Harker, David (2013). How to Split a Theory: Defending Selective Realism and Convergence without Proximity. British Journal for the Philosophy of Science 64 (1):79-106.

Hoefer, Carl (1996). The metaphysics of space-time substantivalism. Journal of Philosophy 93 (1):5-27.

Holland, P. (2005). Computing the Wavefunction from Trajectories: Particle and Wave Pictures in Quantum Mechanics and their Relation. Annals of Physics 315, 505–531

Holland, P. (2017). The quantum state as spatial displacement. In R. Kastner, J. Jeknić-Dugić, and G. Jaroszkiewicz (Eds), Quantum Structural Studies: Classical Emergence from the Quantum Level, pp. 333–72. London: World Scientific Publishing.

Horowitz, G. and Polchinski, J. (2006). Gauge/gravity duality. In D. Oriti (ed.), To wards Quantum Gravity?, Cambridge: Cambridge University Press. arxiv: gr-qc/0602037

Hudetz, Laurenz (2019). Definable categorical equivalence. Philosophy of Science 86 (1):47-75.

Huggett, N. (2017). Target space≠space. Studies in History and Philosophy of Modern Physics, 59, pp. 81-88. http://philsci-archive.pitt.edu/11638/

Huggett, Nick & Wüthrich, Christian (2018). The (A)temporal Emergence of Spacetime. Philosophy of Science 85:1190-1203.

Ismael, Jenann & van Fraassen, Bas C. (2002). Symmetry as a guide to superfluous theoretical structure. In Katherine Brading & Elena Castellani (eds.), Symmetries in Physics: Philosophical Reflections. New York: Cambridge University Press. pp. 371--92.

Jacobs (unpublished). Symmetries as a guide to the structure of physical quantities. DPhil Thesis

Jacobs, Caspar (2022a). Invariance, intrinsicality and perspicuity. Synthese 200 (2):1-17.

Jacobs, Caspar (2022b). The Nature of a Constant of Nature: the Case of G. Philosophy of Science 90 (4):781-97.

Jacobs, Caspar (2023a). Are Models Our Tools Not Our Masters? Synthese 202 (4):1-21.

Jacobs, Caspar (2023b). The metaphysics of fibre bundles. Studies in History and Philosophy of Science Part A 97 (C):34-43.

Jacobs, Caspar (2024). Comparativist Theories or Conspiracy Theories? Journal of Philosophy 121 (7):365-393.

Keller, John A. (2017). Paraphrase and the Symmetry Objection. Australasian Journal of Philosophy 95 (2):365-378.

Knox, Eleanor (2011). Newton–Cartan theory and teleparallel gravity: The force of a formulation. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 42 (4):264-275.

Knox, Eleanor (2014). Newtonian Spacetime Structure in Light of the Equivalence Principle. British Journal for the Philosophy of Science 65 (4):863-880.

Krantz, David ; Luce, Duncan ; Suppes, Patrick & Tversky, Amos (eds.) (1971). Foundations of Measurement, Vol. I: Additive and Polynomial Representations. New York Academic Press.

Le Bihan, Baptiste & Read, James (2018). Duality and Ontology. Philosophy Compass 13:e12555.

Leifer, M. (2014) “Is the Quantum State Real?” Quanta 3, pp. 67-155

Lewis, D. (1983a). Putnam’s paradox. Australasian Journal of Philosophy 62: 221236; reprinted in his Papers in Metaphysics and Epistemology (1999), Cambridge: Cambridge University Press

Luc, Joanna (2023). Motivationalism vs. interpretationalism about symmetries: some options overlooked in the debate about the relationship between symmetries and physical equivalence. European Journal for Philosophy of Science 13 (3):1-33.

Maddy, Penelope (1992). Indispensability and Practice. Journal of Philosophy 89 (6):275.

Maddy, Penelope (2008). How applied mathematics became pure. Review of Symbolic Logic 1 (1):16-41.

Malament, David B. (1995). Is Newtonian cosmology really inconsistent? Philosophy of Science 62 (4):489-510.

Malament, David B. (2012). Topics in the Foundations of General Relativity and Newtonian Gravitation Theory. Chicago: Chicago University Press.

March, Eleanor (forthcoming). Are Maxwell Gravitation and Newton-Cartan Theory Theoretically Equivalent? British Journal for the Philosophy of Science.

March, Eleanor (forthcoming). Categorical Equivalence and the Kinematics-Dynamics Distinction. British Journal for the Philosophy of Science.

March, Eleanor (unpublished). What is the value in an intrinsic formalism?. https://philsci-archive.pitt.edu/25848/

March, Eleanor, Wolf, William J. and Read, James (2024). On the geometric trinity of gravity, non-relativistic limits, and Maxwell gravitation, Philosophy of Physics 2 (1)

Martens, Niels C. M. (2018). Symmetry-to-(un)reality inferences & explanatory power: The case of the Aharonov-Bohm effect. Unpublished draft

Martens, Niels C. M. (2019). The (un)detectability of absolute Newtonian masses. Synthese 198 (3):2511-2550.

Martens, Niels C. M. (2022). Machian Comparativism about Mass. British Journal for the Philosophy of Science 73 (2):325-349.

Martens, Niels C. M. & Read, James (2020). Sophistry about symmetries? Synthese 199 (1-2):315-344.

Maudlin, Tim (1988). The Essence of Space-Time. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1988:82 - 91.

Maudlin, Tim (1993). Buckets of water and waves of space: Why spacetime is probably a substance. Philosophy of Science 60 (2):183-203.

Maudlin, Tim (1998). Healey on the Aharonov-Bohm effect. Philosophy of Science 65 (2):361-368.

Maudlin, Tim (2007). The metaphysics within physics. New York: Oxford University Press.

Maudlin, Tim (2012). Philosophy of Physics: Space and Time. Princeton University Press.

Maudlin, T. (2018). Ontological clarity via canonical presentation: Electromagnetism and the Aharonov-Bohm effect. Entropy 20, 465.

Maudlin, Tim (2019) Philosophy of Physics Volume II: Quantum Mechanics, Princeton: Princeton University Press

Mceldowney, Paul Anh (2020). On Morita equivalence and interpretability. Review of Symbolic Logic 13 (2):388-415.

McSweeney, Michaela Markham (2016). An Epistemic Account Of Metaphysical Equivalence. Philosophical Perspectives 30 (1):270-293.

Meadows, Toby (2024). Beyond Linguistic Interpretation in Theory Comparison. Review of Symbolic Logic 17 (3):819-859.

Møller-Nielsen, Thomas (2017). Invariance, Interpretation, and Motivation. Philosophy of Science 84 (5):1253-1264.

Muller, F. (1995). The equivalence myth of quantum mechanics —Part I. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 28 (1):35-61.

Mundy, Brent (1987). The metaphysics of quantity. Philosophical Studies 51 (1):29 - 54.

Myrvold, Wayne C. (2019). How could relativity be anything other than physical? Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 67:137-143.

Ney, Alyssa (2021). The World in the Wave Function: A Metaphysics for Quantum Physics. New York, NY, USA: Oxford University Press.

North, Jill (2021). Physics, Structure, and Reality. Oxford: Oxford University Press.

Norton, John D. (1992). A Paradox in Newtonian Gravitation Theory. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1992:412 - 420.

Norton, John (2003). Causation as folk science. Philosophers' Imprint 3:1-22.

Nozick, Robert (2001). Invariances: the structure of the objective world. Cambridge: Harvard University Press.

Pirani, F. A. E. (1970). Noncausal Behavior of Classical Tachyons. Physical review D, Particles and fields 1.12. pp. 3224–3225

Pitts, Brian J. (2006). Absolute objects and counterexamples: Jones–Geroch dust, Torretti constant curvature, tetrad-spinor, and scalar density. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 37 (2):347-371.

Pooley, Oliver (2002). Handedness, parity violation, and the reality of space. In Katherine Brading & Elena Castellani, Symmetries in Physics: Philosophical Reflections. New York: Cambridge University Press. pp. 250--280.

Pooley, Oliver (2006). Points, particles, and structural realism. In Dean Rickles, Steven French & Juha T. Saatsi (eds.), The Structural Foundations of Quantum Gravity. Oxford, GB: Oxford University Press. pp. 83--120.

Pooley, Oliver (2013). Substantivalist and Relationalist Approaches to Spacetime. In Robert W. Batterman (ed.), The Oxford Handbook of Philosophy of Physics. Oxford University Press USA.

Pooley, Oliver & Brown, Harvey R. (2002). Relationalism rehabilitated? I: Classical mechanics. British Journal for the Philosophy of Science 53 (2):183--204.

Putnam, Hilary (1977). Realism and Reason. Proceedings and Addresses of the American Philosophical Association 50 (6):483-498.

Putnam, Hilary (1980). Models and reality. Journal of Symbolic Logic 45 (3):464-482.

Putnam, Hilary (ed.) (1983). Realism and Reason. Philosophical Papers (Vol. 3) New York: Cambridge University Press.

Putnam, Hilary (1987). Truth and Convention: On Davidson's Refutation of Conceptual Relativism. Dialectica 41 (1-2):69--77.

Putnam, Hilary (1989). Representation and Reality. Cambridge: MIT Press.

Putnam, Hilary (1990). Realism with a human face. Cambridge: Harvard University Press. Edited by James Conant.

Putnam, Hilary (2001). Reply to Jennifer Case. Revue Internationale de Philosophie 218(4), p. 431-8

Putnam, Hilary (2012). Philosophy in an Age of Science. Harvard University Press

Quine, Willard Van Orman (1948). On what there is. Review of Metaphysics 2 (5):21-38.

Quine, Willard van Orman (1954a). Carnap and Logical Truth. Synthese 12 (4):350-74.

Quine, W. V. (1954b). The scope and language of science. British Journal for the Philosophy of Science 8 (29):1-17.

Read, James (2022). Geometric Objects and Perspectivalism. In James Read & Nicholas J. Teh, The Philosophy and Physics of Noether's Theorems. Cambridge University Press. pp. 257-273.

Read, James (2025). Good VIBES only. MS

Read, James & Møller-Nielsen, Thomas (2020a). Motivating dualities. Synthese 197 (1):263-291.

Read, James & Møller-Nielsen, Thomas (2020b). Redundant epistemic symmetries. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 70:88-97.

Reichenbach, Hans (1958). The philosophy of space & time. New York,: Dover Publications.

Rickles, D. (2011). A philosopher looks at string dualities. Studies in History and Philosophy of Modern Physics, 42, 54–67

Rivat, Sébastien (2021). Effective theories and infinite idealizations: a challenge for scientific realism. Synthese 198 (12):12107-12136.

Rosenstock, Sarita & Weatherall, James Owen (2016). A Categorical Equivalence between Generalized Holonomy Maps on a Connected Manifold and Principal Connections on Bundles over that Manifold. Journal of Mathematical Physics 57:102902.

Rosenstock, Sarita ; Barrett, Thomas William & Weatherall, James Owen (2015). On Einstein Algebras and Relativistic Spacetimes. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 52 (Part B):309-316.

Ruetsche, Laura (2011). Interpreting Quantum Theories: The Art of the Possible. Oxford, GB: OUP

Saunders, Simon (2002). Physics and Leibniz's principles. In Katherine Brading & Elena Castellani, Symmetries in Physics: Philosophical Reflections. New York: Cambridge University Press. pp. 289--307.

Saunders, Simon (2003). Indiscernibles, General Covariance, and Other Symmetries: The Case for Non-Reductive Relationalism. In A. Ashtekar (ed.), Revisiting the Foundations of Relativistic Physics. pp. 151--173.

Saunders, Simon (2007). Mirroring as an a priori symmetry. Philosophy of Science 74 (4):452-480.

Saunders, Simon (2013). Rethinking Newton’s Principia. Philosophy of Science 80 (1):22-48.

Schrödinger, Erwin (1925) Ann. Phys., Lpz 77 325 (Engl. Transl. in Mach’s Principle: From Newton’s Bucket to Quantum Gravity ed. J Barbour and H Pfister (Boston: Birkhäuser))

Schurz, Gerhard (2011). Structural correspondence, indirect reference, and partial truth: phlogiston theory and Newtonian mechanics. Synthese 180 (2):103-120.

Sider, Theodore (2020). The Tools of Metaphysics and the Metaphysics of Science. Oxford, England and New York, NY, USA: Oxford University Press.

Sklar, Lawrence (1982). Saving the Noumena. Philosophical Topics 13 (1):89-110.

Sloan, David & Gryb, Sean (2021). When scale is surplus. Synthese 199 (5-6):14769-14820.

Teitel, Trevor (2021). What Theoretical Equivalence Could Not Be. Philosophical Studies 178 (12):4119-4149.

Trautman, Andrzej (1965). Foundations and current problems of general relativity. Deser, S. & Ford, K. (Eds.), Lectures on general relativity. Englewood Cliffs, N.J.,: Prentice-Hall.

Van Fraassen Bas, C. (1980). The scientific image. New York: Oxford University Press.

Van Fraassen, Bas C. (1989). Laws and symmetry. New York: Oxford University Press.

van Fraassen, Bas C. (1991), Quantum Mechanics: An Empiricist View, Oxford: OUP

Van Fraassen, Bas C. (2008). Scientific Representation: Paradoxes of Perspective. Oxford, GB: Oxford University Press UK.

Wallace, David (2008). The quantum measurement problem: State of play. In Dean Rickles, The Ashgate Companion to Contemporary Philosophy of Physics. Ashgate.

Wallace, David (2012). The Emergent Multiverse: Quantum Theory According to the Everett Interpretation. Oxford, GB: Oxford University Press.

Wallace, David (2019). Who's afraid of coordinate systems? An essay on representation of spacetime structure. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 67:125-136.

Wallace, David (2020). Fundamental and Emergent Geometry in Newtonian Physics. British Journal for the Philosophy of Science 71 (1):1-32.

Wallace, David (2022a). Stating structural realism: mathematics-first approaches to physics and metaphysics. Philosophical Perspectives 36 (1):345-378.

Wallace, David (2022b). Observability, redundancy and modality for dynamical symmetry transformations, in James Read, Bryan Roberts and Nic Teh (eds.), *The Philosophy and Physics of Noether's Theorems: A Centenary Volume* (CUP, 2022), 322-353

Wallace, David & Timpson, Christopher Gordon (2010). Quantum Mechanics on Spacetime I: Spacetime State Realism. British Journal for the Philosophy of Science 61 (4):697-727.

Weatherall, James Owen (2016a). Understanding Gauge. Philosophy of Science 83 (5):1039-1049.

Weatherall, James Owen (2016b). Are Newtonian Gravitation and Geometrized Newtonian Gravitation Theoretically Equivalent? Erkenntnis 81 (5):1073-1091.

Weatherall, James Owen (2016c). Fiber bundles, Yang–Mills theory, and general relativity. Synthese 193 (8).

Weatherall, James Owen (2016d). Regarding the 'Hole Argument'. British Journal for the Philosophy of Science:axw012.

Weatherall, James Owen (2018a). A brief comment on Maxwell[-Huygens] spacetime. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 63:34-38.

Weatherall, James Owen (2018b) Erratum to Weatherall [2016a]. *Philosophy of Science*. 85(2):325-325.

Weatherall, James Owen (2019). Part 1: Theoretical equivalence in physics. Philosophy Compass 14 (5):e12592.

Weatherall, James Owen (2019). Part 2: Theoretical equivalence in physics. Philosophy Compass 14 (5):e12591.

Weatherall, James Owen (2020). Equivalence and Duality in Electromagnetism. Philosophy of Science 87 (5):1172-1183.

Weatherall, James Owen (2021). Why Not Categorical Equivalence? In Judit Madarász & Gergely Székely, Hajnal Andréka and István Németi on Unity of Science: From Computing to Relativity Theory Through Algebraic Logic. Springer. pp. 427-451.

Williams, Porter (2019). Scientific Realism Made Effective. British Journal for the Philosophy of Science 70 (1):209-237.

Williamson, Timothy (2007). The Philosophy of Philosophy. Malden, MA: Wiley-Blackwell.

Williamson, Timothy (2016). Modal science. Canadian Journal of Philosophy 46 (4-5):453-492.

Williamson, Timothy (2017). Modality as a Subject for Science. Res Philosophica 94 (3):415-436.

Wilson, Mark (1981). The double standard in ontology. Philosophical Studies 39 (4):409 - 427.

Wolf, William J., James Read, and Quentin Vigneron (2023). "The Non-Relativistic Geometric Trinity of Gravity," arXiv: 2308.07100

Wolff, J. E. (2020). The Metaphysics of Quantities. Oxford: Oxford University Press

Wu, Jingyi & Weatherall, James (forthcoming). Between a Stone and a Hausdorff Space. British Journal for the Philosophy of Science.